University of Alberta

Cognitively-Active Speaker Normalization Based on Formant-Frequency Scaling Estimation

by

Santiago Barreda Castañón

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

Fall 2013
Edmonton, Alberta

Abstract

The acoustic characteristics associated with a vowel category may vary greatly when produced by different speakers. Despite this variation, human listeners are typically able to identify vowel sounds with a good degree of accuracy. One approach to this issue is that listeners interpret vowel sounds relative to what might be expected for a given speaker, a theory known as *speaker normalization*. This thesis comprises three experiments meant to test specific aspects of a theory of speaker normalization that is under active cognitive-control on the part of the listener, where the information used by the process is organized around the detection of speaker changes. The first experiment investigates the role of f0 in vowel perception, with results indicating that f0 primarily affects vowel quality by influencing the listener's expectations regarding the speaker. In the second experiment, the interaction between the detection of speaker changes and the perception of vowel quality is investigated. Findings support the notion that the detection of speaker changes is a central component of speaker normalization, and that speaker normalization is a cognitively-active process. In the third experiment, listeners were trained to report the acoustic correlate associated with increases or decreases to the average formant frequencies produced by a voice (i.e., formant-frequency scaling). Results indicate that listeners are able to identify voices that differ on the basis of this parameter with good accuracy, and that the perceptual correlate of formant-frequency scaling is influenced by the fundamental frequency of vowel sounds. Finally, a model of cognitively-active speaker normalization, the Active Sliding Template Model (ASTM), is introduced. The ASTM predicts

vowel quality on the basis of a speaker-specific representation that is refined in the absence of a detected speaker change, and re-estimated when a speaker change is detected. An implementation of this model was used to simulate the results of Experiments 1 and 2. The results of these simulations indicate that this relatively simple model of cognitively-active speaker normalization is able to generate a range of patterns of results similar to those observed for human listeners.

Acknowledgements

I would like to thank the faculty and support staff at the Department of Linguistics of the University of Alberta, and in particular Terry Nearey, for not retiring (yet), and for helping me understand what he had been telling me all along.

Table of Contents

List of Tables v
List of Figures vii
Chapter 1 1
1.1 Approaches to resolving ambiguity in vowel sounds 2
1.2 Direct and indirect effects 6
1.2.1 Fundamental frequency and the frame of reference 6
1.2.2 The role of higher formants 7
1.2.3 Some implications of indirect sources of normalization 7
1.3 Active vs. Passive normalization theories 8
1.4 Active speaker normalization 10
1.4.1 The Probabilistic Sliding Template Model of vowel perception 11
1.4.2 Amendments to the Sliding Template Model 13
1.5 Motivation for current work 15
Works Cited 17
Chapter 2 20
2.1 Introduction 20
2.1.1 Direct F0 theories 21
2.1.2 Indirect F0 theories 22
2.1.3 f0-free theories 23
2.1.4 Rationale for the present study 23
2.2 Methodology 27
2.2.1 Participants 27
2.2.2 Stimuli 27
2.2.2.1 F1 and F2 Values 28
2.2.2.2 F3 and higher formants 29
2.2.2.3 Fundamental frequency 29
2.2.2.4 Synthesis of stimuli 30

2.2.3 Procedure 30
2.3 Results 34
2.3.1 Partial correlation analysis 35
2.4 Assessment of the indirectness of effects 38
2.5 General Discussion 42
Works Cited 46
Chapter 3 50
3.1 Introduction 50
3.1.1 Contextual Tuning Theory 52
3.1.1.1 An elaboration of the contextual tuning approach 53
3.1.1.2 Differential predictions of alternative accounts 56
3.1.1.3 Testing Contextual Tuning Theory in Magnuson & Nusbaum (2007) 58
3.1.2 Rationale for current experiment 60
3.2 Methodology 61
3.2.1 Participants 61
3.2.2 Stimuli 62
3.2.3 Procedure 63
3.3 Results 64
3.3.1 Vowel Identification Performance 65
3.3.2 Improvement within a block 67
3.3.3 The relationship between reaction times, hit rates and the detection of speaker changes 70
3.4 Discussion 72
3.5 Conclusion 75
Works Cited 77
Chapter 4 80
4.1 Introduction 80

4.1.1 FF-scaling and apparent speaker characteristics 81
4.1.2 FF-scaling, normalization and vowel perception 84
4.1.3 FF-scaling, vowel perception and apparent speaker characteristics. 87
4.1.3 Rationale for the Current Experiment 89
4.2 Methodology 92
4.2.1 Participants 92
4.2.2 Stimuli 92
4.2.3 Procedure 94
4.3 Results 97
4.3.1 Identification of voice f0 and FF-scaling 98
4.3.1.1 Performance for the Two-factors Task 98
4.3.1.2 Performance for the FF-scaling only Task 99
4.3.2 Information used in FF-scaling estimation 100
4.4 Discussion 103
4.5 Conclusion 108
Works Cited 110
Chapter 5 114
5.1 Summary of results 114
5.1.1 Experiment 1 114
5.1.2 Experiment 2 116
5.1.3 Experiment 3 118
5.2 An explicit model of Active Speaker Normalization 119
5.2.1 The Sliding Template Model 120
5.2.1.1 Control structure implied by the Sliding Template Model 125
5.2.2 The Active Sliding Template Model 126
5.2.2.1 Summary of tuned parameters. 127
5.2.2.2 Detection of speaker changes 128

5.2.2.3 Assessing the appropriateness of current FF-scaling using the refinement and speaker change distance thresholds 129
5.2.2.3 Assessing stability of current FF-scaling using the stability parameter..... 131
5.2.2.4 Updating Extrinsic FF-scaling Estimate 132
5.2.2.6 Finding distances and selecting the winner..... 132
5.3 Simulation of results using the Active Sliding Template Model..... 133
5.3.1 Experiment 1 133
5.3.1.1 Effects for f0 and the higher formants on vowel quality 134
5.3.1.2 Reduction of strength of indirect intrinsic effects..... 137
5.3.2 Experiment 2 139
5.3.2.1 Hit Rates..... 139
5.3.2.2 Reaction Times 141
5.4 Conclusion 144
Works Cited 145
Appendix 1..... 147
Works Cited 151
Appendix 2..... 152
A2.1 Number of voices per block..... 153
B. Confidence in Number of Voices per Block 154
C. Summary of Results 155
Appendix 3..... 156
Works Cited 159

List of Tables

Table 2.1. Formant frequencies and f0s (Hz) used in the creation of the stimuli. 28

Table 2.2. Expected relationships between pairs of variables, all other things being equal. Where appropriate, the intermediate inference leading to this relationship is given. 34

Table 2.3. Results of t-tests performed on the within-participant partial correlation coefficients for pairs of variables. Variables included are F1, F3+, f0, Vowel openness (VO), Maleness (M) and Speaker Size (S). 36

Table 2.4. Mean partial correlation coefficients across all 19 participants for the fully-controlled and no-speaker models. The percent decrease in mean indicates the decrease in magnitude from the fully-controlled model to the no-speaker model as a function of the magnitude of the no-speaker model. 41

Table 3.1. Formant frequencies for the vowels of the baseline voice. 62

Table 4.1. Initial f0 levels for all conditions. Formant frequencies provided are those used for the lowest FF-scaling step vowels in all conditions, corresponding to formant frequencies appropriate for a typical adult male. 93

Table 4.3. Sum of squares and percent of variance explained of judged FF-scaling explained by stimulus FF-scaling (FF-S), stimulus f0 (f0) and the interaction of the two. 102

Table 5.1. Reference patterns specifying the expected F1 F2 and F3 frequencies (in normalized log-Hz) for the vowel phonemes of Edmonton English. If an FF-scaling is added to these values and the sum is exponentiated, the FFs (in Hz) expected for each vowel category given the FF-scaling estimate may be found. 124

Table 5.2. Pooled within-groups covariance matrix given to the ASTM to be used for the classification of vowel sounds. 124

Table 5.3. A summary of tuned parameters involved in the Active Sliding Template Model. 128

Table 5.4. Mean partial correlation coefficients across all 19 participants for the fully-controlled and no-speaker models observed for Experiment 1 (originally presented as Table 2.4) are compared to simulated partial correlation coefficients. The percent change in mean indicates the change in magnitude from the fully-controlled model to the no-speaker model as a function of the magnitude of the no-speaker model. 138

Table A1.1. Percentage of individual vowels (within each speaker group) from the individual data of Hillenbrand et al. (1995) that have f0 values exceeding the frequencies used in the current experiment. 148

Figure A1.2. Distribution of speaker size responses grouped by the vowel's gender response and that vowel's f0. Note that each panel has a different y-axis range. 150

Table A2.1. Percent of rounds in which listeners reported hearing more than one voice in a block, presented by voice-pair type. Numbers in parentheses are the standard errors of each mean..... 153

Table A2.2. Percent of rounds in which listeners reported being unsure of the number of voices in a block, presented by voice-pair type. Numbers in parentheses are the standard errors of each mean. 154

List of Figures

Figure 1.1. The average vowel system of adult males (left panel) and children (right panel) from the Peterson and Barney (1952) data. The line indicates the points where ln F1 = ln F2. The point on the figure represents a vowel sound with an F1 of 580 Hz and an F2 frequency of 1220 Hz. 2

Figure 1.2. In the left panel, three possible reference spaces are compared, where each is represented by a polygon. The point on the figure indicates the absolute location of a vowel sound with an F1 of 580 Hz and an F2 frequency of 1220 Hz. In the right panel, each symbol represents the relative location of the vowel sound from the left panel with respect to the reference spaces in the right panel, where the symbol used to show relative locations indicate which reference space they are based on. 3

Figure 1.3. In each panel, the black point indicates the location of a given vowel sound, while the circle, cross and triangle represent three candidate vowel categories. The dotted lines show examples of possible movement in the reference-space estimates. If the template is allowed to move in any direction, as in the left panel, the vowel may be made to match any category. If movement is limited to movement along lines parallel to ln F1 = ln F2, the number of candidate vowel categories will be limited. 11

Figure 2.1. Screenshot of the experimental interface..... 32

Figure 2.2. Distributions (across participants) of average partial correlation coefficients between pairs of variables (V.O. = vowel openness, S.S. = Speaker Size, Male = Maleness). The dotted lines represent bounds at which an individual participant's coefficient reaches significance ($p < 0.05$) 38

Figure 2.3. Partial correlation coefficients (averaged over participants) between pairs of variables (V.O. = Vowel Openness, S.S. = Speaker Size, Male = Maleness). The broken line between Size and Vowel Openness indicates the only relationship which did not reach significance by t-test. Arrows indicate the presumed direction of effects. 39

Figure 3.1. Average within-participant hit rate, presented by voice-pair type. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean. 66

Figure 3.2. Average coefficient relating within-block target number, and expected hit rates for that target within a block. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean. 69

Figure 3.3. Average, within-participant residual resulting from regressing reaction time on hit rates, presented by voice-pair type. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean. 71

Figure 4.1. In the left panel, ellipses enclosing two standard deviations of the Peterson and Barney (1952) vowels are presented. Vowels have been normalized-using the log-mean normalization method of Nearey (1978). F1 and F2 are presented as a the ratio of each formant frequency to the geometric mean F1-F2-F3 frequency produced by each speaker across their whole vowel system. The line is a line parallel to ln F1 = ln F2. In the right panel, all formant frequencies have been log-transformed and centered within-category so that vowel-category means are at the origin. All points have been rotated 45 degrees clockwise so that the ln F1 = ln F2 line is now parallel to the x-axis (Dimension 1), while Dimension 2 represents the orthogonal axis. The lines indicate the major axes of the vowel-category ellipses, which no longer vary primarily along the ln F1 = ln F2 axis (Dimension 1)..... 86

Figure 4.2. The x axis indicates the mean of the first three formant frequencies for productions of /i/. Ellipses enclose two standard deviations of the distribution of real voices from data collected by Hillenbrand et al. (1995). Ellipses indicate the distribution of voices of adult males (dotted line),

adult females (solid line) and children (broken line). The locations of stimulus voices at each ΔFF–scale level are indicated by the filled points. 93

Figure 5.1. Scatterplot of speakers from two data sets (Peterson and Barney 1952; Hillenbrand et al. 1995) plotted according to their FF-scaling (indexed by log-mean FFs) and log-mean f0. The bold line indicates the regression line predicting log f0 on the basis of FF-scaling. The dotted lines parallel to the regression line indicate one standard deviation in f0 given the FF-scaling. 122

Figure 5.2. Histogram of the marginal (prior) distribution of FF-scalings from across two data sets (Peterson and Barney, 1952; Hillenbrand et al. 1995). The bold line shows the density of a normal distribution with the same mean and variance parameters as this marginal distribution. 122

Figure 5.3. A flowchart representing the control structure implied by the Sliding Template Model as described in Nearey and Assmann (2007). 125

Figure 5.4. A flowchart representing the processes of the Active Sliding Template Model. The stages shared by the unmodified Sliding Template Model are shaded in grey. The letters in brackets indicate where the parameters outlined in Table 5.3 are used by the model..... 126

Table 5.3. A summary of tuned parameters involved in the Active Sliding Template Model. 128

Figure 5.5. A comparison of observed and simulated percentage of /æ/ responses for the data from Experiment 1. Data is pooled across continuum steps, within F3+ and f0 condition. Letters indicate F3+ (first letter) and f0 (second letter) condition from among: Low (L), Medium (M) and High (H). The solid line indicates points along which x = y. To the extent that the simulation accurately reflects listener behaviour, points should all fall along this line. The dotted line indicates points along which simulated /æ/ responses are 6.4% greater than observed responses. 135

Figure 5.6. The percent /æ/ responses, organized by f0 level, observed for Experiment 1 are indicated by the solid line. The dotted line shows

classification patterns of the ASTM when the conditional variance of f0 given FF-scaling is set equal to the value suggested by Nearey and Assmann (2007). The broken line shows classification patterns by the same model when this parameter is divided in half. 136

Figure 5.7. Hit rates are compared for observed results, and those predicted by the full Active Sliding Template Model (Simulation A). The solid line indicates blocks where voices had dissimilar source characteristics, the dotted line indicates blocks where voices had dissimilar source characteristics. 140

Figure 5.8. Hit rates are compared for two modified versions of the Active Sliding Template Model. The solid line indicates blocks in which voices had dissimilar source characteristics, the dotted line indicates blocks where voices had dissimilar source characteristics. 140

Figure 5.9. Response times observed for participants in Experiment 2 are compared to CPU times for different voice-pair types. In both cases only times for hit rates are reported. CPU times are in nanoseconds estimated with reference to the CPU clock. Solid lines indicate blocks where voiced had the same source characteristics while the broken line indicates blocks where these differed. 143

Figure A1.1. Kernel density plots for the f0 measurements in the data of Hillenbrand et al. 1995. The vertical lines represent the three f0 levels used in the current experiment. 148

Figure A1.2. Distribution of speaker size responses grouped by the vowel's gender response and that vowel's f0. Note that each panel has a different y-axis range. 150

Figure A3.1. In the left panel, ellipses enclosing two standard deviations of the Peterson and Barney (1952) vowels are presented. The line is a line parallel to ln F1 = ln F2. In the right panel, all formant frequencies have been log-transformed and centered within-category so that vowel-category means are at the origin. All points have been rotated 45 degrees clockwise so that the ln F1 = ln F2 line is now parallel to the x-axis (Dimension 1),

while Dimension 2 represents the orthogonal axis. The lines indicate the major axes of the vowel-category ellipses, and they all vary around the ln F1 = ln F2 axis (Dimension 1). 157

Chapter 1

Introduction

Perceived vowel quality is most strongly determined by the frequencies of the first two formants of a vowel sound (Joos 1948; Peterson 1961; Nearey 1978; Rakerd and Verbrugge 1985; Miller 1989). For decades, the most common summary of the acoustic properties of vowels has been based on a plot of the frequencies of the first two formants, as seen in Figure 1.1. This simple representation serves as a useful vehicle to introduce the problem of speaker differences, and an early approach to the problem of vowel normalization that can be labeled in its most general form as a *frame of reference* approach.

Researchers familiar with formant-range differences between adult males and children would not be surprised to find patterns like those seen in Figure 1.1. Although the absolute formant frequencies associated with a certain vowel category across the two speaker classes are different, the relative positions of the vowels are very similar in the two panels. This has led several researchers to suggest that the first two formant frequencies of a vowel sound (and its position within the 2-formant space) are interpreted relative to a speaker-specific frame of reference, rather than in an absolute manner.

Various alternative accounts of vowel perception, notably those involving higher-dimensional representations of vowel sounds or cognitively-passive auditory processes will be considered later in this chapter, and throughout the body of this thesis. However, it will be argued that a general frame of reference approach to vowel perception, driven by cognitively-active control structures, best accounts for the results obtained in the series of experiments to be outlined here, and for a range of experimental results reported in previous works.

1.1 Approaches to resolving ambiguity in vowel sounds

One of the earliest approaches to resolving ambiguity in the formant space were theories involving a speaker specific 'frame of reference' used to interpret the vowels of different speakers (Joos, 1948; Ladefoged and Broadbent, 1957; Nearey, 1978). Joos (1948) suggested that the vowels of different speakers may be "phonetically identical, although acoustically distinct" as long as "each of them occupies the same position within the vowel quadrilateral of the speaker" (p. 59). Essentially, the idea is that vowel quality is not determined on the basis of the absolute acoustic characteristics of the sound, but on these characteristics judged relative to what should be expected from the speaker.

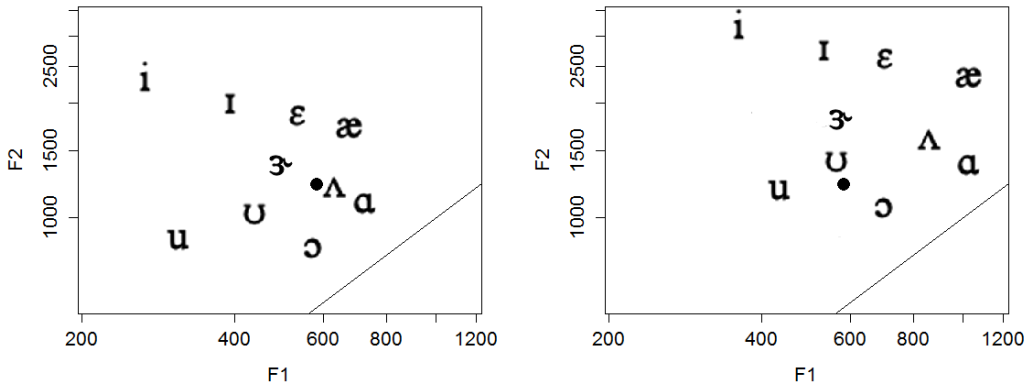

Figure 1.1. *The average vowel system of adult males (left panel) and children (right panel) from the Peterson and Barney (1952) data. The line indicates the points where ln F1 = ln F2. The point on the figure represents a vowel sound with an F1 of 580 Hz and an F2 frequency of 1220 Hz.*

This may be visualized with the aid of Figure 1.1. A single point has been indicated in each panel with a black circle. When considered relative to the vowel system in the left panel, this point would seem to be an instance of /ʌ/. On the other hand, relative to the system in the right panel, a vowel at the same location within the 2-formant space is more likely to be an instance of /ʊ/.

The 'frame of reference' may be represented by the expected location of each vowel phoneme within the formant space, given the speaker (Nearey, 1978). Traditional frame of reference theories focused on a 2-formant space, representing vowel sounds based solely on the frequencies of the first two formants.

Consequently, for these theories, the frame of reference can be thought of as a subsection of the F1-F2 plane as seen in Figure 1.2. Nearey (1978) compared this to a 'sliding template' that the listener moves around based on expectations regarding the speaker. Once the location of this template has been set, vowel quality may be determined based on the relative location of the vowel sound on the template. This frame of reference, consisting of the expected locations of vowel categories within the formant-space for a given speaker, will be referred to as the reference space[1].

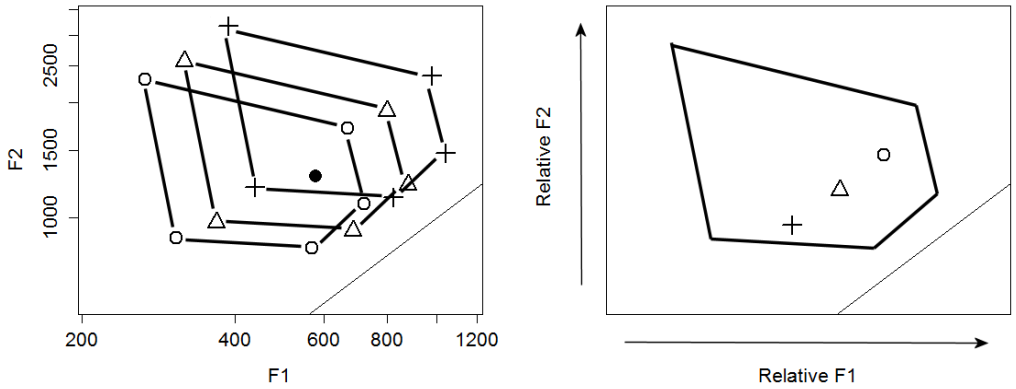

Figure 1.2. *In the left panel, three possible reference spaces are compared, where each is represented by a polygon. The point on the figure indicates the absolute location of a vowel sound with an F1 of 580 Hz and an F2 frequency of 1220 Hz. In the right panel, each symbol represents the relative location of the vowel sound from the left panel with respect to the reference spaces in the right panel, where the symbol used to show relative locations indicate which reference space they are based on.*

For example, the point in the left panel of Figure 1.2 has a single absolute location within the 2-formant space. However, as seen in the right panel of the same figure, this vowel may have several different relative locations when compared to different reference spaces. This approach to vowel normalization suggests that committing to an interpretation of a vowel sound necessarily

[1] For the purposes of statistical testing or pattern classification, the frame of reference can also be represented as an assortment of vectors, where each row or column indicates the expected FFs for a vowel category. This vector would specify the location of the vowel category in the formant space.

involves committing to a particular frame of reference used to interpret the vowel. Because of their reliance on a speaker-specific frame of reference, theories of this kind will be referred to as *speaker normalization* theories.

The early versions of speaker normalization theories involved an extrinsically-specified frame of reference which required that the listener have some amount of accumulated evidence before speech produced by the speaker could be correctly identified. Joos (1948) provides the following account: "On first meeting a person, the listener hears a few vowel phones, and on the basis of this small but apparently sufficient evidence he swiftly constructs a fairly complete background (coordinate system) upon which he correctly locates new phones as fast as he hears them" (p. 61). Following Ainsworth (1975), and Nearey (1989), this view of normalization is referred to as *pure-extrinsic* because establishing the frame of reference requires that a listener gather information from one (or more) previous speech sounds.

This account of vowel perception can explain experimental results such as those presented in Ladefoged and Broadbent (1957), which showed that the spectral characteristics of a precursor phrase can affect the classification of following target sounds in a manner consistent with predictions made by theories of speaker normalization. However, they cannot explain how listeners are able to correctly identify isolated vowels and syllables as accurately as they are, in some cases above 90% correct even when 10 or more candidate vowel categories are being considered (Assmann et al. 1982). Furthermore, F3 and f0 have been found to affect perceived vowel quality in several previous studies (Fujisaki and Kawashima 1968; Nearey 1989; Slawson 1968), suggesting that a 2-formant representation of vowel sounds may be inadequate.

In part as a reaction to these issues, some researchers proposed intrinsic models of vowel specification, which focused on determining higher-dimensional representations of vowel sounds based only on the intrinsic (internal) properties of vowel sounds (Miller 1989; Sussman 1986; Syrdal and Gopal 1986). For example, Syrdal and Gopal (1986) emphasize that their normalization method is an "inherent or speaker-independent normalization procedure. Only the acoustic

parameters present in an individual segment are used in the normalization. In contrast, speaker-dependent normalizations require a sampling of vowels from the same speaker" (p. 1095).

These intrinsic (or speaker independent) vowel perception theories sought to eliminate or reduce between-category overlap in the formant space by organizing vowels within a space with three or more dimensions. Usually these dimensions involve the first three formant frequencies (FFs) and, optionally, f0. These approaches typically represent vowel sounds based on the ratio of the frequencies of adjacent formants, or based on the ratio of the FFs to some normalizing value (e.g., mean FFs or f0). Although pure-intrinsic models can help explain how listeners are able to identify single speech tokens from a speaker, they have the inverse weakness of pure-extrinsic models in that they cannot explain why extrinsic factors can affect perceived vowel quality. Furthermore, models which necessarily involve f0 and/or F3 in the specification of vowel quality also have difficulties explaining experimental results that show that listeners can identify whispered or two-formant vowels (Nusbaum and Morin 1992; Eklund and Traunmüller 1997).

Nearey (1989) suggested that both intrinsic and extrinsic information may play a role in the perception of vowel sounds, and that neither pure-intrinsic nor pure-extrinsic models of vowel perception could accurately account for listener behaviour in all listening situations. Following this, I will adopt the position that vowel quality is determined by a process of speaker normalization whereby the reference space (or frame of reference) estimate is determined on the basis of both intrinsic and extrinsic information. Another way to look at this is that evidence gleaned from the current speech sound is considered jointly with information taken from previous speech sounds.

However, the process of combining extrinsic and intrinsic information provides the listener with additional problems not typically considered by traditional views of speech perception. For example, it seems reasonable to think that listeners do not accumulate extrinsic evidence from all prior speech events and use them to interpret all future speech events regardless of the appropriateness

of the prior extrinsic information, especially since extrinsic information is only useful to the extent that it leads to an appropriate reference-space estimate. This suggests that for extrinsic information to be optimally useful there should be some mechanism which ensures that a listener uses it only when it is beneficial to do so.

1.2 Direct and indirect effects

Although there is general agreement that the first two formants are directly involved in the specification of vowel sounds, there is less agreement on the role of f0 and the higher formants (formants higher than F2). Johnson (1990a) distinguished between direct and indirect effects on vowel quality. A cue with a direct effect on vowel quality is involved in its specification, as is likely to be the case for the first two formant frequencies of a vowel sound. In contrast, indirect effects influence perceived vowel quality by affecting the reference space used to interpret them. The general proposed mechanism is that indirect effects provide the listener with information regarding the type of speaker they are hearing, and so provide information regarding likely expected formant-patterns. For example, consider the two possible interpretations of a vowel sound presented in Figure 1.1. Children generally produce higher FFs overall than male speakers. Any acoustic cues associated with children (e.g., a high f0) may suggest to the listener that the speaker produces higher FFs overall, and so might influence the listener to adopt the formant space shown in the right panel as the working frame of reference, thereby affecting perceived vowel quality.

1.2.1 Fundamental frequency and the frame of reference

Previous studies have found that vowel quality shifts can be induced by manipulating extrinsic or intrinsic f0 (Miller 1953; Fujisaki and Kawashima 1968; Slawson 1968; Nearey 1989; Johnson 1990a). Johnson (1990a, 1999, 2005) has suggested that f0 affects vowel quality primarily indirectly, by affecting the reference-space estimate rather than by being directly involved in the specification of vowel quality. For example, Johnson (1990a) found that vowel quality shifts induced by f0 were minimized in situations where listeners were unlikely to associate changes in f0 with speaker changes. Furthermore, the effect of f0 on

perceived vowel quality has been found to be sensitive to mode of presentation (Johnson, 1990a; Nusbaum and Magnuson, 1992) and instructions (Magnuson and Nusbaum, 2007). Finally, to my knowledge there is no vowel quality that relies on a specific f0 in the way that some vowels depend on F3, for example, as rhotic vowels necessarily have low F3 values. As a result, following Johnson (1990a), f0 will be considered to have a primarily indirect effect on vowel quality.

1.2.2 The role of higher formants

Although, F3 has been found to affect perceived vowel quality in previous experiments (Fujisaki and Kawashimi 1968; Slawson 1968), its effect is typically much weaker than that of the first two FFs. Nearey (1989) has suggested that F3 may offer the listener with intrinsic extrinsic-information. That is, because of its weaker relationship to vowel quality and stronger relation to vocal-tract length, F3 may offer information that is (nearly) directly relatable to the reference space without having to listen to a range of speech from that speaker. However, F3 has been demonstrated to be crucial to the perception of front vowels (Fujimura 1967; Johnson 1989), and is crucial to the perception of rhotic vowels. As a result, it seems fair to include F3 in the direct specification of vowel quality.

1.2.3 Some implications of indirect sources of normalization

The notion of effects that indirectly affect perceived vowel quality leads to some interesting possibilities regarding theories of vowel perception. If a cue were to affect vowel quality only indirectly, we would expect that it should have no effect on vowel quality in situations in which it did not affect the reference space. For example, consider two vowel sounds with the same FFs but an octave difference in their respective mean f0s. If f0 has only an indirect effect and a listener assumes each vowel was produced by a different speaker, they may use the intrinsic f0 for each vowel to estimate two appropriate frames of reference. This might lead to quite different estimates and to two different perceived vowel qualities. On the other hand, if the listener assumes that both vowels are produced by the same speaker, the change in f0 may be disregarded and a single frame of

reference may be used to interpret both vowels, leading to a single perceived vowel quality regardless of the differences in f0.

Essentially, the strength of indirect effects may vary based on the outcome of processes that mediate the relationship between indirect effects and perceived vowel quality. As with the issue of correct use and segregation of extrinsic information, the regulation of indirect effects would benefit from variable behaviour based on the expectations of the listener.

1.3 Active vs. Passive normalization theories

Nusbaum and Magnuson (1997) distinguish between three components of the process of speaker normalization: representations of information, transformations of representations, and control structures that regulate the transformations to be carried out (i.e., the input-output mapping). When viewed in this way, pure-intrinsic theories of vowel perception seek to explain vowel normalization only with reference to the representation of vowel sounds, while speaker normalization theories do the same by focusing on transformations of these representations. However, the question regarding the kind of control structure that governs vowel normalization is also an important one.

Nusbaum and Magnuson (1997) distinguish between passive (open-loop) control structures and active (closed-loop) control structures. Passive control structures lead to processes that exhibit a deterministic, predictable mapping between input and output. In contrast, active control structures involve non-deterministic mappings between input and output. Furthermore, the relationship between inputs and outputs may be modified in a context-sensitive manner, and based on the results of the output of the system.

Nusbaum and colleagues (Nusbaum and Morin 1992; Nusbaum and Magnuson 1997; Magnuson and Nusbaum 2007) suggest that vowel normalization is under *active* cognitive control, and outline two types of evidence in support of cognitively-active normalization. The first of these is that active processes are expected to result in increased demands on cognitive processes and working memory. Listeners exhibit just these sorts of processing costs when presented with mixed-speaker listening conditions. Listeners have been found to

identify words more slowly (Summerfield and Haggard 1973) and less accurately (Creelman 1957) in mixed versus blocked speaker conditions. For example, Nusbaum and Morin (1992) asked participants to remember a series of numbers during a speech identification task and found that this increased reaction times only in mixed-speaker conditions. Wong et al. (2004) report the results of an fMRI study in which they found that participants listening to mixed-speaker lists showed a greater degree of activation in the middle/superior temporal and superior parietal regions of the brain, indicating that mixed-voice listening conditions can result in different and increased processing.

The second type of evidence supporting active normalization is processing flexibility of the kind not easily achieved by cognitively-passive control mechanisms. For example, the negative effect on performance associated with accounting for differences between speakers has been found to vary based on listener expectations (Johnson et al. 1999; Johnson 1990b; Magnuson and Nusbaum 2007). Magnuson and Nusbaum (2007) presented listeners with lists of synthetic voices which differed only slightly in their mean f0s. One group was instructed that the list contained only a single voice, while the other group was told the list was composed of multiple voices. The authors found an increase in response times only for the group that was told to expect multiple voices in the block. This suggests that processes associated with speaker normalization are not automatically carried out every time a listener encounters a new speech sound, and that there is some level of control over whether to behave as if a listening situation contains multiple speakers.

Nusbaum and colleagues suggest that speaker normalization is guided by a process they refer to as *contextual tuning*. According to contextual tuning, the listener uses whatever information is available to them to arrive at an appropriate mapping between the acoustic signal and a listener-internal representation of speech sounds. This representation may be refined as new evidence becomes available and is discarded if a change in speaker is detected. This approach to vowel perception can help to resolve two of the major issues associated with speaker normalization theories highlighted previously. First, since the listener is

hypothesized to actively monitor for speaker changes, the incorrect use of extrinsic information is minimized. Extrinsic information (including a priori assumptions induced by, for example, instructions or photographs of possible speakers) may be used in the absence of a detected speaker change, while it may be disregarded in the presence of a detected speaker change. Secondly, information which indirectly affects perceived vowel quality may have a strong effect on perceived vowel quality in cases in which it is associated with a detected speaker change, and a weak effect, or no effect at all, in cases where it is not.

1.4 Active speaker normalization

The processes that compose contextual tuning theory are generally similar to those suggested by previous researchers, in particular, those that posit that vowel perception occurs on the basis of a speaker-dependent frame of reference. For example, the very idea that vowels be compared relative to a speaker-specific coordinate system outlined in Joos (1948) presupposes that there be some system that monitors for a change in speaker, lest extrinsic information accumulate over a listener's entire lifetime. Weenink (2006) has proposed several speaker-adaptive normalization methods which update the frame of reference based on the output of the state of the current system. These models are able to replicate the perceptual advantage seen in single-voice listening conditions over mixed-voice conditions. The important difference is that in contextual tuning, changes and refinements of the frame of reference are explicitly organized according to detected speaker changes. Furthermore, the detection of speaker changes, and the refinement of the frame of referenced are hypothesized to be under active-cognitive control.

Contextual tuning is a theory regarding the *control structure* of vowel normalization, which does not make strong claims about the nature of the other two components of vowel perception (representations and mappings of representations). In the remainder of this section, I will briefly outline an explicit model of vowel perception that may be used to fill in the blanks, and create a more complete, and testable, theory.

1.4.1 The Probabilistic Sliding Template Model of vowel perception

Theories of speaker normalization suggest that listeners interpret vowel sounds relative to the sounds that might be expected for a given speaker, rather than in an absolute manner. To the extent that these theories are correct, one of the greatest problems facing the listener is estimating a likely reference-space for an unknown speaker. This can be visualized as choosing from among each of the reference spaces shown in the left panel of Figure 1.2 to use as a speaker-dependent frame of reference.

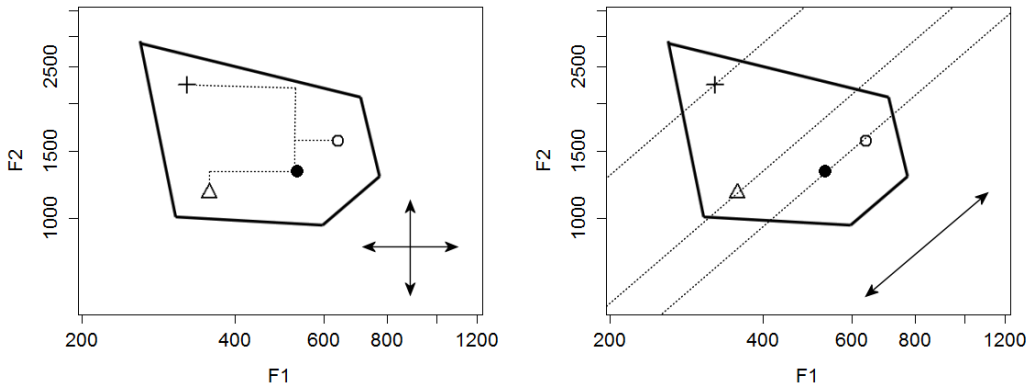

Figure 1.3. *In each panel, the black point indicates the location of a given vowel sound, while the circle, cross and triangle represent three candidate vowel categories. The dotted lines show examples of possible movement in the reference-space estimates. If the template is allowed to move in any direction, as in the left panel, the vowel may be made to match any category. If movement is limited to movement along lines parallel to ln F1 = ln F2, the number of candidate vowel categories will be limited.*

If the reference spaces of different speakers of the same dialect could differ in arbitrary ways, estimating an appropriate reference space for a given listener might prove very difficult, if not impossible. For example, in the left panel of Figure 1.3, if one were to move the template to the left and down, the right, the point could be made to look like an instance of a back-high vowel. On the other hand, if one moved the template down and to the left, the same vowel would appear to be a low-central vowel.

Fortunately for listeners, there is good evidence that the formant patterns produced by different speakers of the same dialect differ from each other by a

single multiplicative parameter[2] (Nearey and Assmann 2007; Nearey 1978; Nordström and Lindblom 1975; Turner et al. 2009). Essentially, despite anatomical differences between individuals, speakers strive to produce formant patterns which differ from those of other speakers of their language by a single scalar value when producing a given vowel category.

As noted by Nearey (1978), variation according to a single multiplicative parameter means that the reference spaces of speakers of the same language differ from each other along lines parallel to ln F1 = ln F2, where *ln* represents the natural logarithm function. This situation is represented in the right panel of Figure 1.3. If the template is only allowed to move along lines parallel to ln F1 = ln F2, the relative locations of vowel sounds are also limited to vary along lines parallel to ln F1 = ln F2, greatly reducing the plausible interpretations of any given speech sound. For example, in the right panel of Figure 1.3, only the circle is a plausible interpretation for the indicated point. Furthermore, if reference spaces vary according to a single parameter, the exact location of the reference space (and therefore the expected locations of individual vowel categories) could be indexed with reference to this parameter, indicating its location along the ln F1 = ln F2 axis (Nearey 1978).

This suggests that the process of vowel normalization could be considered to depend on the estimation of this parameter for a given speaker (Nearey 1978; Nearey and Assmann 2007). I will refer to this parameter as the formant-frequency scaling, or FF-scaling, of a speaker, where speakers with higher FF-scalings produce higher formant frequencies overall. Given an FF-scaling estimate, and knowledge of the relative locations of vowel categories in the formant space, the expected formant patterns representing the vowels produced by a certain speaker can be estimated. As noted by Nearey (1978), this may be visualized as sliding the reference template in the right panel of Figure 1.3 along the tracks created by the dotted lines, where higher FF-scaling estimates lead to

[2] This issue is dealt with in depth in the introduction of Chapter 4 of this thesis, and in Appendix 3.

template positions closer to the upper right corner of the Figure. Once an appropriate template location (and associated FF-scaling estimate) has been determined, the most likely vowel category to generate that observed formant pattern may be predicted (Nearey and Assmann 2007).

Nearey and Assmann (2007) outline several explicit models of vowel perception which classify vowels using a single, speaker-specific parameter (which they refer to as Ψ, analogous to FF-scaling) and a language-specific reference template indicating expected formant-patterns for the vowels of the language. Each model predicts perceived vowel quality in the same way, by taking the FF-scaling estimated for a trial, and modifying the observed FFs to compare them to the reference template. The authors describe several models which differ in the ways they estimate FF-scaling or in the manner that they specify the template vowel patterns.

Of particular interest is Method 6, which estimates a speaker's FF-scaling using only the intrinsic information carried in vowel sounds. Method 6 estimates a speaker-specific FF-scaling parameter based on the covariance of f0 and FF-scaling across human speakers and the fit of the given vowel pattern to each candidate vowel category. Consequently, Method 6 can be classified as a method of intrinsic, indirect-f0 normalization. Method 6 of the sliding template model offers a very useful explicit model of human speech perception which follows the general speaker normalization approach to vowel perception, and incorporates the apparent indirect effect of f0 on perceived vowel quality. However, this model can be amended slightly to incorporate some of the insights of contextual tuning theory in order to make an even more complete model of human vowel perception.

1.4.2 Amendments to the Sliding Template Model

There are three accommodations that need to be made to Method 6 of the Sliding Template Model in order to make it compatible with contextual tuning

theory[3]. The first of these is to allow for variable cue weights based on the expectations of by the listener. This would allow, for example, for the strength of the effect for f0 on vowel quality to change as a result of listener expectations.

The second modification is to allow for more variable forms of information to affect speaker-dependent FF-scaling estimates. For example, vowel quality shifts have been induced by manipulating apparent speaker gender independently of any acoustic characteristics (Glidden and Assmann, 2004; Johnson et al., 1999). This suggests that the model could benefit from some mechanism to allow salient apparent speaker characteristics, such as speaker gender, to indirectly affect perceived vowel quality by affecting the speaker-dependent FF-scaling estimate.

The final modification is to give the process a memory, which can combine the current output with previous outputs when appropriate, and which discards previous estimates when not appropriate. Although the exact mechanisms by which this might occur are not at all clear, the general way this might occur may be sketched out. On first hearing a speaker, the listener uses intrinsic information to arrive at an FF-scaling estimate in a manner similar to that outlined by Method 6 of the Sliding Template Model. In the absence of perceived speaker changes, this representation may be refined as necessary by combining new intrinsic information with previous extrinsic information. In situations where a speaker change is detected, the current estimate may be discarded and a new estimate may be formed based solely on the intrinsic properties of the new speech sound. Finally, the process of FF-scaling estimation may itself help resolve issues related to the detection of speaker changes in that gross mismatches between the current reference-space and the new formant pattern may be good evidence of a likely change in speaker.

---

[3] Nearey and Assmann (2007) mention mechanisms by which each of the modifications to be suggested may be made to the sliding template model. Generally, these are carried out by affecting the relative weights associated with the cues of the current stimuli, or the weights associated with the current information relative to the prior expectations of the system. See Section 5.2 for more information.

1.5 Motivation for current work

The literature reviewed above suggests that, although the specific details regarding speaker normalization are unknown, the broad outlines are fairly clear. In general, experimental evidence supports a theory of speaker normalization in which vowel quality is determined relative to expectations regarding the reference space of the current speaker. To the extent that the reference spaces of different speakers of the same dialect differ by a single parameter (i.e., FF-scaling), the process of speaker normalization could be considered to center around the estimation of this parameter. The estimation of this parameter is speaker specific, and the use of extrinsic information and the strength of indirect effects are both governed by a cognitively-active process that monitors changes in speaker.

The chapters that make up the body of this thesis present the results of three individual experiments that were carried out in order to investigate specific aspects of speaker normalization based on FF-scaling estimation. Chapter 2 presents the results of an experiment designed to test the directness of the effect of f0 on perceived vowel quality. The findings of this experiment suggest that this effect is primarily indirect, and that f0 affects perceived vowel quality mostly by having a strong influence on apparent speaker characteristics. These may influence FF-scaling estimates, and thence judgments of vowel quality mediated by expectations regarding the speaker.

Chapter 3 reports an experiment carried out to investigate the extent of active-cognitive control over the process of FF-scaling estimation, by investigating identification performance for different kinds of mixed-voice listening conditions (conditions which feature multiple voices presented at random). The results of this experiment indicate that identification performance for a mixed-speaker list cannot solely be explained by mismatches in the appropriate reference spaces of the speakers involved. Rather, evidence is presented which suggests that identification accuracy for a mixed-voice listening condition is best explained when the facility with which listeners can identify speaker changes is also taken into account.

In Chapter 4, an experiment investigating the ability of listeners to report FF-scaling directly is described. The results of this experiment shows that listeners can learn to respond to FF-scaling with good accuracy after only a short training session, and suggest that there may be a perceptual quality which is primarily related to voice FF-scaling. However, some results also indicate that the perceptual correlate of FF-scaling may be influenced by f0, so that perceived FF-scaling may not solely dependent on the range of FFs associated with a given voice.

Each of these chapters is presented with its own introduction and conclusion sections that, in some cases, frame the issues at hand in a very general way. This was done intentionally in order that the experimental results would have a broad appeal and be applicable to as many theories of vowel perception as possible, as is appropriate for a journal article[4]. For example, the indirect effect of f0 on vowel quality reported in Chapter 2 should be taken into account by theories of vowel perception regardless of whether or not a researcher agrees with the argumentation laid out in the earlier sections of this chapter regarding speaker normalization relying on a single scale-factor. However, in point of fact, all these experiments were conceived and designed with the intent of refining and investigating specific aspects of the theory outlined in the previous sections, regardless of their more general presentation in the following chapters. The relation of specific results to this theoretical perspective will be further discussed in Chapter 5.

[4] The body of this thesis (chapters 2-4) were published as journal articles in Barreda and Nearey (2012), Barreda (2012), and Barreda and Nearey (2013) respectively.

Works Cited

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant and M. Tatham (Eds.). Auditory analysis and perception of speech. London: Academic Press. 103-113.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America*, *71*(4), 975–989. doi:10.1121/1.387579

Barreda, S. and Nearey, T. (2013). Training listeners to report the acoustic correlate of vocal tract length using synthetic voices. Journal of the Acoustical Society of America 133(2): 1065-1077.

Barreda, S. (2012). Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. Journal of the Acoustical Society of America 132(5): 3453-3464. [Link]

Barreda, S. & T.M. Nearey. (2012). The direct and indirect roles of fundamental frequency in vowel perception. Journal of the Acoustical Society of America 131: 466-477.

Creelman, C. D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, *29*, 655. doi:10.1121/1.1909003

Fujimura, O. (1967). On the Second Spectral Peak of Front Vowels: a Perceptual Study of the Role of the Second and Third Formants. *Language and Speech*, *10*(3), 181–193. doi:10.1177/002383096701000304

Fujisaki, H., and Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 73– 77. doi:10.1109/TAU.1968.1161952

Glidden, C. M., and Assmann, P. F. (2004). Effects of visual gender and frequency shifts on vowel category judgments. *Acoustics Research Letters Online*, *5*(4), 132–138. doi:10.1121/1.1764472

Johnson, K. (1989). Higher formant normalization results from auditory integration of F2 and F3. *Perception and psychophysics*, *46*(2), 174–180.

Johnson, Keith. (1990a). The role of perceived speaker identity in F0 normalization of vowels. *The Journal of the Acoustical Society of America*, *88*(2), 642–654. doi:10.1121/1.399767

Johnson, Keith. (1990b). Contrast and normalization in vowel perception. *Journal of Phonetics*, *18*(229), 54.

Johnson, Keith, Strand, E. A., and D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. doi:10.1006/jpho.1999.0100

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Ladefoged, P., and Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. doi:10.1121/1.1908694

Magnuson, J. S., and Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. doi:10.1037/0096-1523.33.2.391

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America*, *85*(5), 2114–2134. doi:10.1121/1.397862

Miller, R. L. (1953). Auditory Tests with Synthetic Vowels. *The Journal of the Acoustical Society of America*, *25*(1), 114–121. doi:10.1121/1.1906983

Nearey, T. M. (1978). Phonetic Feature Systems for Vowels. PhD thesis, Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America 85. 2088–2113.

Nearey, T. M. and Assmann, P. F. (2007). Probabilistic "sliding template" models for indirect vowel normalization. In Maria-Josep Solé, Patrice Beddor, and Manjari Ohala (eds.) Experimental Approaches to Phonology. Oxford: Oxford University Press. 246-69.

Nordström, P.-E., and Lindblom, B. (1975). *A normalization procedure for vowel formant data*. University.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 175-184.

Peterson, G. E. (1961). Parameters of Vowel Quality. *Journal of Speech and Hearing Research*, *4*(1), 10.

Rakerd, B., and Verbrugge, R. R. (1985). Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels. *The Journal of the Acoustical Society of America*, *77*(1), 296–301. doi:10.1121/1.392393

Slawson, A. W. (1968). Vowel Quality and Musical Timbre as Functions of Spectrum Envelope and Fundamental Frequency. *The Journal of the Acoustical Society of America*, *43*(1), 87–101. doi:10.1121/1.1910769

Summerfield, Q., and Haggard, M. P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report of speech research in progress*, *2*, 12–23.

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language*, *28*(1), 12–23. doi:10.1016/0093-934X(86)90087-8

Syrdal, A. K., and Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100. doi:10.1121/1.393381

Turner, R. E., Walters, T. C., Monaghan, J. J. M., and Patterson, R. D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *The Journal of the Acoustical Society of America*, *125*(4), 2374. doi:10.1121/1.3079772

Weenink, D. J. M. (2006). *Speaker-adaptive vowel identification*. SpeechMinded. Retrieved from http://dare.uva.nl/document/37721

Wong, P. C. M., Nusbaum, H. C., and Small, S. L. (2004). Neural Bases of Talker Normalization. *Journal of Cognitive Neuroscience*, *16*(7), 1173–1184. doi:10.1162/0898929041920522

Chapter 2

The direct and indirect roles of fundamental frequency in vowel perception

2.1 Introduction

Listeners are able to recognize the vowels of their language with relative ease despite the fact that the physical characteristics of these sounds can vary a good deal from speaker to speaker. However, the same variation that may hinder speech perception affords the listener with a wealth of information regarding the speaker. For example, listeners are able to judge the gender of an adult speaker with relative ease (Bachorowski and Owren 1999; Strand 2000; Perry et al. 2001). They are also able to make consistent judgements regarding the apparent size of the speaker using only information available from that speaker's voice[5] (van Dommelen and Moxness 1995; Lass et al. 1980; Collins 2000; Smith and Patterson 2005; Smith et al. 2005; Rendall et al. 2007). If, and how, apparent-speaker characteristics and phonetic information interact in the determination of speech sounds is an open question in speech perception.

[5] There is a general correlation between speaker size and average vocal tract length (Fitch and Giedd 1999, Hollien et al. 1994) across genders and speakers of all ages and sizes. However, after controlling for age and gender, there is no correlation between speaker height and weight and estimated vocal tract length (van Dommelen and Moxness 1995, Collins 2000, Gonzalez 2004). There is also no correlation between speaker height and weight and average speaking f0 when controlling for gender and age (Lass and Brown 1978, Kunzel 1989). As a result, it is not surprising that several studies have found that listeners are not very good at judging the actual size of a speaker solely on the basis of their speech (van Dommelen 1993, Collins 2000, Rendell 2007). Although listeners are not very accurate when estimating speaker size, their estimates, both correct and incorrect, have been found to be fairly consistent both within and between listeners (van Dommelen and Moxness 1995, Lass et al. 1980, Collins 2000), and are strongly influenced by both f0 and the FFs (Collins 2000, Smith and Patterson 2005, Smith et al. 2005, Rendall et al. 2007).

From the perspective of speech production, the fundamental frequency (f0) and formant frequencies (FFs) of a vowel are more or less independent (Fant 1960) so that f0 should have only a small effect on the spectral content of a vowel. If vowel quality were entirely determined by the spectral content of a vowel, a change in f0 alone should cause no change in vowel quality. However, many experiments have induced vowel quality changes by changing intrinsic and/or extrinsic f0 with respect to the FFs of a vowel (Miller 1953, Fujisaki and Kawashima 1968, Slawson 1968, Johnson 1990, Glidden and Assmann 2004). Studies have induced similar effects by changing only the expected gender of the speaker. This has been done by presenting alternating male and female faces with identical stimuli (Glidden and Assmann 2004) and by telling listeners to imagine either a male or female speaker (Johnson et al. 1999). It is not clear why changes in apparent-speaker characteristics and changes in f0 affect vowel quality or if they do so independently or via the same general mechanism.

There are three general schools of thought regarding the relationship between vowel quality, f0 and apparent-speaker characteristics. These will be referred to as direct f0 theories, indirect f0 theories and f0-free theories.

2.1.1 Direct F0 theories

Average f0 tends to co-vary with average FFs across speakers (Hollien 1994; Fitch and Giedd 1999; Nearey and Assmann 2007). In general, larger people have lower FFs and f0s while smaller people have higher FFs and f0s. Listeners have been found to show a sensitivity to this covariance. They rate speech as more natural (Assmann and Nearey 2007) and identify vowels correctly at a higher rate (Lehiste and Meltzer 1972; Gottfried and Chew 1986; Assmann and Nearey 2008) when speech has the expected relationship between f0 and FFs. For these and related reasons, some researchers suggest that listeners take advantage of this covariance and, as a result, that f0 is directly related to vowel quality in the same way the FFs are (Syrdal and Gopal 1986; Miller 1989). These theories will be referred to as direct f0 theories. The net effect of these theories is empirically indistinguishable from one in which f0 is used by listeners as a scaling

factor to eliminate inter-speaker differences by interpreting FFs in relation to f0 (Nearey 1989, 1992).

2.1.2 Indirect F0 theories

Others, such as Nearey and Assmann (2007) and Johnson (1990, 1999, 2005), suggest that f0 is most important in determining certain apparent-speaker characteristics rather than in the specification of vowel quality directly. According to these theories, f0 is related to vowel quality only insofar as it contributes to the determination of whichever apparent-speaker characteristics affect vowel quality. These theories will be referred to as indirect f0 theories. Johnson (1990) suggests that listeners create a mental representation of the speaker and that speech is interpreted on the basis of the characteristics of this presumed speaker. In this model, f0 is only used to determine likely speaker identity. Johnson (2005) takes this several steps further and outlines an exemplar based *Talker Normalization* model:

> "Rather than warp the input signal to match a fixed internal template, the internal representation adapts according to the 'perceived identity of the talker' (Johnson 1990), as exemplars appropriate for the talker are activated and inappropriate exemplars are deactivated. [...] cues of all kinds can be involved in tuning the activated set of exemplars [... including] F0 as a gender cue" (p. 383)

Other researchers suggest that indirect normalization takes place via more abstract apparent-speaker characteristics rather than properties tied to limited classes of exemplars. For example, the Probabilistic Sliding Template Model (PSTM) (Nearey and Assmann 2007) works on the basis of Ψ, which is a speaker-dependent value roughly equivalent to the average FF produced by a speaker. By adding Ψ*, an estimate of Ψ, to a language-specific reference pattern, a listener can estimate expected FFs for the vowels of that language as produced by a given speaker. The PSTM uses f0, as well as information about the distribution of average FFs and the relationship between FFs and f0 to estimate the most likely Ψ for that speaker. (See also Traunmüller (1994) for a rather more elaborate account

of an indirect relationship between observed f0 of a specific stimulus and perceived vowel quality; this approach may make predictions similar to those of the indirect normalization theories considered above, at least in some circumstances).

2.1.3 f0-free theories

A final possibility is that there is no relationship between f0 and vowel quality. These theories will be referred to as f0-free theories. Despite the results of experiments reported in Section 2.1.2 above, Patterson and colleagues have made strong claims about the independence of f0 and vowel quality. In a series of experiments that manipulate spectrum envelope and f0 independently via a vocoder, they found that changes in f0 have virtually no effect on vowel quality[6] (Smith et al. 2005).

To explain this, Smith et al. (2005) and Irino and Patterson (2002) have suggested that the auditory system performs a Mellin(-like) transform on the acoustic input at an early stage in auditory processing. This results in a *size-shape image* (Irino and Patterson 2002, 188) in which the spectral pattern of a sound is represented as an invariant shape and the size of the resonator that produced the sound is represented as the position along one dimension of the sound pattern in the sound-shape image. In this view, changes in f0 or in apparent-speaker properties play no role in determining vowel quality.

2.1.4 Rationale for the present study

All three of the above theories could be considered different forms of vowel normalization, where normalization refers to a process by which a listener removes or compensates for speaker-specific variation from an incoming vowel token. We are treating the normalization process as a black box where we may

[6] However, Smith et al.'s experiment used only five phonetically dissimilar seed vowels /i, e, a, o, u/ from a single speaker. In experiments using similar vocoding techniques, but 12 vowel categories and several speakers, Assmann and Nearey (2008) found considerable variation in vowel identification rates as a function of the relation between spectrum-envelope scaling and f0.

observe the input (the physical properties of the stimuli) and the output (vowel quality) but not the internal workings of the system. We do not seek here to determine the exact internal workings of the normalization process, but simply to consider what kinds of information may affect the transfer characteristics of the process.

The experiment to be described in the following pages was designed to test the relationship between f0, vowel quality and apparent-speaker characteristics. To do this, a vowel continuum was matched with several different f0s and higher formants (in this case, formants higher than F2 which will be referred to as F3+). The general stimulus design is similar to that of Fujisaki and Kawashima (1968) and to the isolation condition in experiments described in Johnson (1990). In fact, the experiment to be outlined here could be viewed as an extension and refinement of some of the experiments described in Johnson (1990). Because of the importance of some of the results presented in that paper to our current experiment, some of the relevant results will be summarized.

Johnson used a series of synthetic /hVd/ tokens with varying formant and f0 levels which were intended to be interpreted as either /ʊ/ or /ʌ/.Vowels were presented in two conditions: an isolation condition and a phrase condition. In the isolation condition, vowels were presented in a random order (with no extrinsic context) so that the intrinsic f0 of a vowel stimulus varied randomly from trial to trial. This would have resulted in something like a 'speaker-randomized' condition. In the phrase condition, the same /hVd/ stimuli were presented following a synthetic voice saying "This is" which had either a rising intonation (simulating a question) or a falling intonation (simulating a declarative).

Johnson conducted an AX-discrimination pretest using stimuli with a single set of formant frequencies, but many f0 levels. Listeners were presented with pairs of stimuli and asked to judge whether the two syllables were spoken by the same or different speaker. Results indicated that although two tokens with the same f0 might be very likely to be judged as being from the same speaker in the isolation condition, the opposite is the case in the phrase condition since a speaker is unlikely to use the same final f0 for a phrase with falling and rising intonation.

Johnson also conducted a second pretest, where listeners provided judgments of speaker size and gender for the stimuli of the AX pretest. The results provide evidence that size judgments are affected by the likelihood of perceived speaker differences as measured in the AX test.

Based on the results of the AX pretest, Johnson designed three vowel classification experiments involving a seven-member formant continuum and two f0 levels per experiment. These experiments were intended to test the relationship between apparent speaker changes and vowel perception. These experiments and the AX pretest were carried out with different groups of participants. Using this methodology, Johnson found an association between the likelihood of a perceived change in speaker in the pretest and the magnitude of an f0-induced vowel category shift in the main experiments. In listening conditions in which listeners were likely to hear different speakers, f0-induced shifts were maximized. In conditions in which listeners were likely to hear a single speaker these same effects were minimized. This association applied to both the isolated word and the phrasal presentation conditions.

Johnson presents a strong circumstantial case for the relationship between f0-induced vowel quality changes and apparent-speaker characteristics. Although his conclusions rely on some very reasonable inferences, they are inferences nonetheless. Specifically, the methodology does not allow for insight into the decisions listeners make on a trial-by-trial basis; nor, for that matter, does it allow for insight into the behaviour of any one listener in both the pretests and the main experiments, since different listeners were involved in all cases.

The experiment to be described below represents, in a sense, an amalgamation of aspects of both pretests and of the isolation conditions of Experiments 1 and 2 of Johnson (1990). For each stimulus presented, we asked participants to make simultaneous judgments of vowel quality and two aspects of speaker characteristics, so that analysis could proceed on a token-by-token basis. Johnson found large effects of f0 for isolated syllables in Experiment 1 where f0 and formant patterns varied from trial to trial. When more information is available about an apparent speaker's intonation and (possibly) formant ranges, the effect of

f0 on vowel quality may be greatly reduced[7]. Our experiment uses isolated vowels with complete randomization of all stimulus properties from trial to trial, resulting in what amounts to a speaker-randomized condition with little to no extrinsic context.

By simultaneously collecting both vowel quality information and apparent-speaker characteristics, we can relate f0-induced vowel quality shifts to changes in the apparent speaker. Although we are not asking for listeners to identify speaker changes directly, the collection of speaker gender and size information will allow us to control for important aspects of perceived speaker changes from the perspective of the listener at the moment of the vowel judgment.

If f0 and apparent-speaker characteristics do not contribute to the determination of vowel quality, they should not have a significant relationship to vowel quality after the formant frequencies have been accounted for. If f0 is directly related to vowel quality, there should be a stable and consistent relationship between f0, the FFs, and vowel quality. Additionally, after these physical properties have been taken into account, there should be no relationship between vowel quality and apparent-speaker characteristics. If f0 affects vowel quality mainly indirectly via its effect on apparent-speaker characteristics, there could be a variable and complicated relationship among judgments of apparent-speaker characteristics, f0 and vowel quality. Furthermore, the relationship between f0 and vowel quality should be considerably weaker, or perhaps non-existent, once apparent-speaker characteristics are controlled for.

[7] Other sources of variation, such as vocal effort, may also affect the relation between stimulus properties and perceived vowel quality (See Traunmüller 1994). However, for monosyllablic stimuli in mixed-speaker type presentation with a simple falling intonation pattern, it seems unlikely that these potential sources of variance would have much effect. Furthermore, any effect they did have would simply tend to weaken any relations of the kind we are studying here and should not add any spurious correlations.

2.2 Methodology

2.2.1 Participants

Listeners were 19 students from the University of Alberta, 16 females and 3 males drawn from a participant pool in which undergraduate students take part in experiments in exchange for partial course credit. They ranged in age from 17 to 54 years old. All were students taking an introductory level, undergraduate linguistics course.

2.2.2 Stimuli

The vowel continuum was constructed on the basis of naturally produced data collected from Edmonton English speakers. A continuum was designed that spanned from roughly the average F1-F2 frequencies of the /ʌ/ of a male to those of the average /æ/ produced by a female in seven equal logarithmic steps. The vowels used were chosen because, when produced by Western Canadian English speakers, they fall on a line almost exactly parallel to the line F1 = F2 in log-formant space. This meant that a single scale factor could be applied to both formants to either change vowel identity or to approximate the change in FFs because of a change in speaker size. Additionally, production data collected at the Alberta Phonetics Laboratory indicated that F3 was nearly identical for the two vowels, meaning that it carried little to no phonetic information. As a result of this F3 could be manipulated without greatly affecting the phonetic quality of the vowels, at least for vowel stimuli consistent with those of a single speaker. The low F3 level was set using perceptual data also collected at the Alberta Phonetics lab. An F3 frequency was selected at which the /ʌ/-/æ/ boundary was perpendicular to the F1 = F2 line so that F1 and F2 would contribute about equally to possible category boundary shifts.

The fourth point of this continuum had F1-F2 frequencies appropriate for either an /æ/ produced by an adult male or an /ʌ/ produced by an adult female. This seven-step continuum was combined with three different F3+ conditions and three different f0 conditions for a total of 63 different vowels. The stimuli were designed in a log space using ln (Hz) (the natural logarithm of the frequency in

Hz). The frequencies of all of the continuum points and f0 and F3 levels used are presented in Table 2.1.

	f0 Levels		
	Low	**Mid.**	**High**
Initial	120	170	240
Final	96	136	190

F3 Levels		
Low	**Mid.**	**High**
2475	2755	3068

Step #	**1**	**2**	**3**	**4**	**5**	**6**	**7**
F1	684	735	789	848	911	978	1051
F2	1354	1455	1563	1679	1803	1937	2081

Table 2.1. *Formant frequencies and f0s (Hz) used in the creation of the stimuli.*

2.2.2.1 F1 and F2 Values

Since the vowels fall almost exactly parallel to the F1 = F2 line in log space, F1 and F2 were modified at the same rate and are therefore perfectly correlated. For this reason they will be treated as one variable, which for the sake of brevity will simply be referred to as F1. The formants for each successive step were about 0.0713 natural log units higher than those of its predecessor. This corresponds to an increase of about 7.4% in Hz. A three step difference in the F1 continuum corresponds to a 0.214 ln (Hz) change (a 22.5% increase). This is about one third the difference between the typical male /ʌ/ (Step 1) and the typical male /æ/ (Step 4) and also the difference between the typical male /æ/ (Step 4) and the typical female /æ/ (Step 7). Therefore, a change of F1 of this magnitude (0.214 ln (Hz), or 22.5%) corresponds to the distance between these phonemes for a single speaker and to the average difference between the phonemes as produced by males and females.

2.2.2.2 F3 and higher formants

The low F3 was set at a value typical for adult males and the highest value was calculated by increasing the log frequency by the previously mentioned male to female step (.214 ln (Hz), or 22.5%). The intermediate F3 value is the (geometric) mean of the high and low F3s. The low F4 was set at 3200 Hz and every successive FF (F5-F11) was set at 1100 Hz higher than the previous FF. The intermediate higher formant frequencies were raised by 11% relative to low higher formants and high higher formants were raised by an additional 11% relative to the intermediate higher formant frequencies. The factor corresponding to F3 and the higher formants will be called F3+.

2.2.2.3 Fundamental frequency

The low f0 level was set to 120 Hz, appropriate for an adult male. The high f0 level was set to reflect the natural covariance between FFs and f0. Nearey and Assmann (2007) report that in a log scale f0 increases 0.31 times as fast as typical FFs, which is close to the value of 1/3 used by Miller (1989) to relate the logs of F1 and f0. This means that, for example, a speaker who produces an average f0 1.0 ln (Hz) higher than a second would also be expected to produce FFs that are 0.31 ln (Hz) higher (roughly 36%), on average, than this second person. In accordance with this relationship, the high f0 condition was one octave[8] higher than the low condition, which we set at a value appropriate for a male speaker. This resulted in a high f0 value of 240 Hz, which was considered appropriate for an adult female. The intermediate f0 condition is the (geometric) mean of the high and low f0s. The f0 values described above refer to the initial f0. The f0 contour decreased linearly across the vowel to a value 0.80 times the initial value.

The f0 levels in this experiment reflect the range observed for adults in Hillenbrand et al. (1995). Specifically, the lowest f0 used was 120 Hz which is about 0.51 standard deviations lower than the average male value (mean = 131

[8] An octave increase in Hz, is equal to 0.693 ln (Hz). This times the 0.31 scale factor gives us 0.214, the difference between the FFs of males and females.

Hz, s.d. = 22 Hz). The highest f0 was about 0.84 standard deviations above the average adult female values observed (mean = 220 Hz, s.d. = 23 Hz).

2.2.2.4 Synthesis of stimuli

All vowels were 225 ms in duration, with steady-state formants. They were synthesized using an implementation of a Klatt (1990) synthesizer provided on version 5.1.01 of Praat (Boersma and Weenink, 2009) and synthesized at a sampling rate of 44.1 kHz. The Praat Klatt synthesizer works on the basis of tiers, each of which contains a separate piece of information about the sound to be synthesized. A single voice source tier was created containing the source specifications to be used for all vowels across all conditions. The source was created with a special focus on the female voice it would create, so that it would sound like a naturally produced female voice and not a male voice with a high pitch. This was accomplished by using a slightly breathy voice source, and small negative spectral tilt, both of which have been found to be associated with femininity in North American English (Price 1989, Klatt and Klatt 1990, Mendoza et al. 1996, Van Borsel et al. 2009).

Three pitch tiers were created, one for each of the three f0 conditions. Tiers were also created containing formant frequency and bandwidth information for the higher formants, formants 3-11, in each of the three F3+ conditions. Because of the high sampling rate, 11 formants were found necessary to fill the Nyquist band and prevent excessive energy roll-off at higher frequencies. Formant bandwidths were set to the larger of 6% of the formant frequency or 60 Hz. All sounds were synthesized using the single voice source and every combination of formant and pitch tier for all three conditions resulting in 9 distinct conditions (3 pitch conditions x 3 formant conditions).

2.2.3 Procedure

Participants were instructed that they would be hearing a human-like, ‘robotic’ voice producing vowels intended to be either /ʌ/ or /æ/. Participants were asked to listen to the vowel and decide which of the two vowel categories the vowel sounded most like. In a pilot experiment, we asked participants to indicate

how tall and how masculine/feminine the speaker they just heard was. We found that masculinity and femininity correlated strongly with f0 and that it may have been too specific a quality. Furthermore, many participants had difficulty reporting the height of the speaker; some were not familiar with the imperial system (we asked for heights in feet and inches) while others felt that height was too specific, they thought the synthetic speakers varied by being more or less muscular or bulky rather than by being taller or shorter. Rather than ask participants for the continuous judgments of masculinity/femininity and height of the speaker, we asked participants for two kinds of judgments about apparent-speaker characteristics:

1. A discrete gender judgment.
2. A graded size judgment, specific definition of size was left for the participants to interpret as they saw fit. The size judgement was intended to correlate with the listener's approximate vocal tract length, and hence formant ranges.

We left the definition of size deliberately vague because of difficulties encountered in pilot experiments that used absolute physical units. The lack of explicit instructions given to participants and the fact that the size scale might have been used in different ways within each gender may have led to differences in how listeners used the size scale. (See Appendix 1). However, any resultant increase in variability would only add noise to the data. It thus seems unlikely to bias any patterns in the data in any specific direction relevant to the hypotheses at hand.

Participants were presented with the sounds over headphones in a sound-attenuated booth and responses were recorded on a computer interface using software specifically designed by the first author for this experiment. Vowel quality responses were input by recording clicks of a mouse on a response button 800 pixels in length, where the x-axis coordinate of the pixel on which the participant clicked was entered as the response so that responses were recorded on an 800 point rating scale. Vowel responses were recorded on a button that said

Hud (corresponding to /ʌ/) on one end and *Had* (corresponding to /æ/) on the other end. Participants were told that the selection of vowel had to fall into one category or the other and that clicking towards the extremes indicated the degree to which the vowel they had just heard sounded more like one vowel than the other. This scale was aligned so that a larger value corresponded to a more /æ/-like vowel. For this reason, this measure will be referred to as the Openness of the vowel.

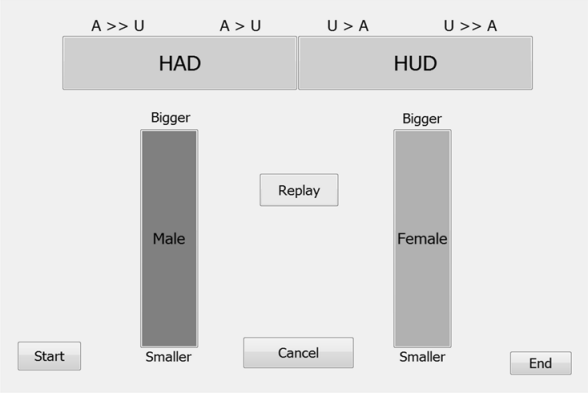

Figure 2.1. *Screenshot of the experimental interface.*

Speaker Size responses were recorded on two separate buttons, one indicating a male speaker and one indicating a female speaker. Participants were instructed that selection of speaker size was also continuous and that clicking higher on the size button indicated a larger speaker. The size/gender buttons were 400 pixels high, in this case the y-axis coordinate at which the participant clicked was entered as the size response. The Speaker-Size judgment scale was aligned so that a larger value corresponded to a larger speaker. Size responses were recorded on two separate buttons, one labelled *male* and the other *female*, which were placed orthogonally to the vowel response button. The use of two separate size buttons, one for each gender allowed us to collect simultaneous gender and size information with a single click. Speaker gender was coded so that a value of 0 corresponded to a female speaker and 1 corresponded to a male speaker. Since

this value indicates a male speaker, this value will be referred to as Maleness. A screenshot of the experimental interface is provided in Figure 2.1. To control for any spurious correlation between vowel and speaker judgments due to horizontal arrangement of the response buttons, the left -right position of the Male and Female response boxes was counter-balanced across listeners.

The procedure was as follows: A stimulus was presented, after which participants had to make a vowel quality and speaker size and gender determination. After these three values had been provided, the next stimulus would play after a 500 ms pause. Vowel sounds were presented in a random order along all stimulus dimensions. Participants were told they could repeat a stimulus up to 2 more times by hitting a button marked *replay* but only if they had not selected any responses for that stimulus. To cancel or undo any selections they had made, participants could click on a button marked *cancel* which erased any answers already provided for the current and previous stimuli, placed them both back into the upcoming stimuli queue, and re-shuffled the queue.

Participants took part in experimental sessions of approximately one hour in length. Before beginning the experiment, participants completed a short training session during which they became familiar with the tasks and the response interface. During the training session participants heard naturally produced /hVd/ syllables containing either /æ/ or /ʌ/ in which the stimuli were produced by two male and two female speakers. Standard practice was to have participants listen to three repetitions of the stimulus list (189 responses), followed by a short break, after which the participant performed another three repetitions of the same list. In some cases, participants were not able to perform all six repetitions of the stimuli list. In these cases, only the data from completed repetitions was used. A total of 6,921 responses were collected across all 19 participants.

Evidence:	Inference:	Expected Effect on variable:
Higher F1		More open vowel
Higher F3+	Shorter vocal tract	Less open vowel
Higher f0	Shorter vocal tract	Less open vowel
Higher formants/f0	Less likely to be male	Female response
Higher formants/f0	Smaller speaker	Lower speaker size response
Larger speaker Size	Longer vocal tract	More open vowel
Male	Longer vocal tract	More open vowel

Table 2.2. *Expected relationships between pairs of variables, all other things being equal. Where appropriate, the intermediate inference leading to this relationship is given.*

2.3 Results

To organize a discussion of the results, we will outline the expected relationships between pairs of variables according to an indirect f0 theory in which f0 changes vowel quality by affecting a listener's frame of reference[9] which in turn is assumed to be correlated with vocal tract length and formant frequency ranges. These relationships correspond to the expected correlations with *all other things being equal*. Open vowels occur with F1 frequencies near a speaker's maximum F1. A speaker with larger vocal tract has a lower maximum F1 than a speaker with a shorter vocal tract. If interpreted as coming from a speaker with a larger vocal tract, a stimulus with an intermediate F1 will appear to be nearer to that speaker's maximum F1 and hence sound more open. As a result, evidence which would lead a listener to conclude that the speaker is larger should lead to the perception of a relatively more open vowel, while evidence to the contrary would result in the perception of a relatively less open vowel for any given set of formant frequencies. A summary of related predictions is presented in Table 2.2, assuming average natural relations between gender, f0 and vocal tract length.

[9] Our usage of the term 'frame of reference' here and in the discussion refers to the formant space that is likely to be used by a speaker. This usage is consistent with the tradition of Joos (1948), Ladefoged and Broadbent (1957) and Nearey (1989).

This experiment contained three manipulated variables (F1, F3+ and f0) and three response variables (vowel openness, Maleness and Speaker Size). The manipulated variables were controlled experimentally and are not affected by any other variables. The response variables are the three variables whose values are provided by the listeners. These reflect properties that exist only in the mind of the listener and may interact with the manipulated variables, and with each other, in unknown ways.

2.3.1 Partial correlation analysis

To investigate the relationship between these variables, a series of within-participant partial correlations was conducted. By considering the partial correlations between pairs of variables after controlling for all of the remaining variables we can investigate the relationship between these variables independently (of any linear effects) of all the others. For example, the partial correlation between f0 and vowel quality after controlling for F1, F3+ and Speaker Size and Maleness will tell us how f0 and vowel quality are expected to co-vary for a vowel with given formant frequencies when produced by a speaker of given apparent size and gender. The process to be outlined below was carried out for each pair of response variables (vowel openness, Maleness and Speaker Size) and every combination of individual response variable and individual manipulated variable (F1, F3+ and f0). The process will be outlined using the relation between f0 and Speaker Size as an example.

The following procedure was applied to the data of each listener in turn. To investigate the relationship between f0 and Speaker Size independently of all of the other variables in the experiment, each of these two variables was regressed in turn on the remaining four variables (F1, F3+, vowel openness, Maleness). After this, the correlation between the residuals from the two regressions was found. The resulting partial correlation coefficient corresponds to the correlation between f0 and Speaker Size after controlling for the effects of all of the remaining variables. In this particular case, it is expected that f0 will be negatively related to Speaker Size since higher f0s should be associated with smaller speakers. If, all other things being equal, participants associate higher f0s with

smaller speakers, then the partial correlation between Speaker Size and f0 should, on average, be significantly different from zero. If participants do not associate smaller speakers with higher f0s then the expected value of average partial correlation between f0 and Speaker Size after controlling for F1, F3+, vowel openness will be zero. Since this correlation is bi-directional, any discussion of cause and effect is dependent on the variables involved. For example, it is presumed that f0 causes the change in vowel openness rather than the other way around, since f0 is controlled by the stimulus design. Causal relations between pairs of judged qualities, however, are indeterminate.

This process was repeated for all 12 pairs of variables considered. This resulted in 19 partial correlation coefficients (one for each listener) for each of the 12 variable pairs. Following the two-stage procedure of Lorch and Myers (1990), independent sample t-tests were performed on the coefficients for every pair of variables to see if the results were significantly different from zero, on average across participants. The results of the t-tests are presented in Table 2.3.

Relation	**F1, VO**	**F3+, VO**	**f0, VO**	**F1, M**	**F3+, M**	**f0, M**
Mean Corr.	0.802	-.215	-.053	-.152	-.147	-.744
t (d.f. 18)	64.5	-10.9	-3.02	-6.49	-9.02	-41.0
p. value	<.001	<.001	.007	<.001	<.001	<.001

Relation	**F1, S**	**F3+, S**	**f0, S**	**M, VO**	**S, VO**	**S, M**
Mean Corr.	-.212	-.151	-.374	.049	.027	-.475
t (d.f. 18)	-9.18	-4.05	-8.59	3.00	1.06	-12.2
p. value	<.001	<.001	<.001	<.008	.303	<.001

Table 2.3. *Results of t-tests performed on the within-participant partial correlation coefficients for pairs of variables. Variables included are F1, F3+, f0, Vowel openness (VO), Maleness (M) and Speaker Size (S).*

All except the last column of Table 2.3 relate directly to patterns predicted by the general indirect-f0 normalization model discussed at the beginning of

Section 2.3 as summarized in Table 2.3. Notably all are in the expected direction and all are significant at $p < .01$ level or better, save for the relationship between Speaker Size and vowel openness[10]. Although the relation between Speaker Size and vowel openness does not reach significance using a t-test, 14 out of 19 listeners show a positive relationship between the two variables, a result that is not likely to have occurred by chance ($p = .022$ via a non-parametric binomial test).

The predictions of Table 2.2 involve relationships between specific stimulus properties and listener judgments of vowel quality or speaker characteristics, or between speaker characteristics and vowel quality. However, the last column of Table 2.3 involves the relation between the judgments of the two apparent-speaker characteristics, controlling for all other factors. The significant negative partial correlation between Speaker Size and Maleness is at first surprising, since one would expect voices heard as Male to be associated with larger absolute sizes. There are, it turns out, reasonable explanations for the negative partial correlation actually observed. These are discussed in Appendix 1.

Figure 2.2 shows the distributions of the coefficients of Table 2.3 across listeners. In the discussion below, references to relative strength of relationship between variable pairs will be based on the average magnitude (absolute value) of the partial correlation coefficient so that a variable pair with a larger magnitude will be deemed to have a stronger relationship than one with a smaller magnitude.

F1 and F3+ both relate strongly to vowel openness, though F1 is a stronger determinant. With the exception of one listener, the distribution of coefficients for the F1 to vowel openness relationship are tightly clustered around the mean, while the coefficients representing the relation between vowel openness and F3+ are more equally distributed over a wider area. The relations between F1 and Maleness and F1 and Speaker Size are only slightly stronger than those between Speaker Size and F3+ and Maleness and F3+. It seems that both F1 and F3+ affect

[10] All variable pairs except Speaker Size and vowel are significant at less than a Šidak adjusted one-tailed (in the expected direction) single test level of $p = 0.00874$ for a family size of 12.

both vowel quality and apparent speaker size, but that F1 is more strongly linked to vowel quality while F3+ is more strongly linked to apparent-speaker characteristics. Maleness is related to all three of the manipulated cues, though f0 is its strongest determinant. Speaker Size is also determined jointly by considering all three manipulated variables and f0 is also its strongest determinant.

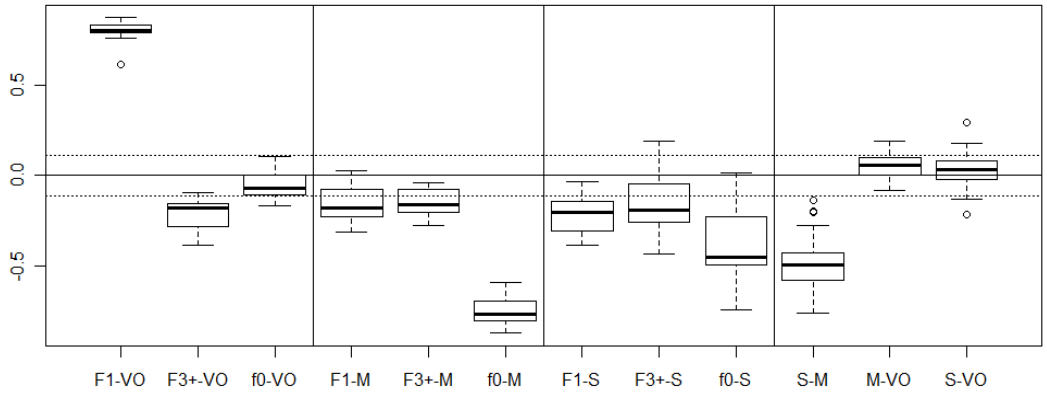

Figure 2.2. *Distributions (across participants) of average partial correlation coefficients between pairs of variables (V.O. = vowel openness, S.S. = Speaker Size, Male = Maleness). The dotted lines represent bounds at which an individual participant's coefficient reaches significance ($p < 0.05$)*

2.4 Assessment of the indirectness of effects

The fact that all of the relations presented in Table 2.3 are in the expected direction (all but one significantly so) is taken as evidence that the basic structure of the design was successful. Since the stimuli were synthesized using parametric synthesis, no real speaker identity or vowel quality can be associated with any of the stimuli other than whatever properties are attributed to the sound or speaker on the part of the listener. However, participants demonstrated an ability to extract both vowel quality and apparent-speaker characteristics from the stimuli. Furthermore, they interpreted this information in a fairly consistent way.

Figure 2.3 presents the same information found in Table 2.3 and Figure 2.2 but in a manner that is easier to inspect visually. The arrows between variables indicate the presumed direction of the effects and the numbers besides each

variable indicate the average strength of the effects. The direct effect of a manipulated variable on the response variables can be judged by the average strength of the direct connection between the two variables. The indirect effect of a manipulated variable can be gauged by considering the effects the variable had on one or more of the response variables jointly with the effects the response variables have on each other.

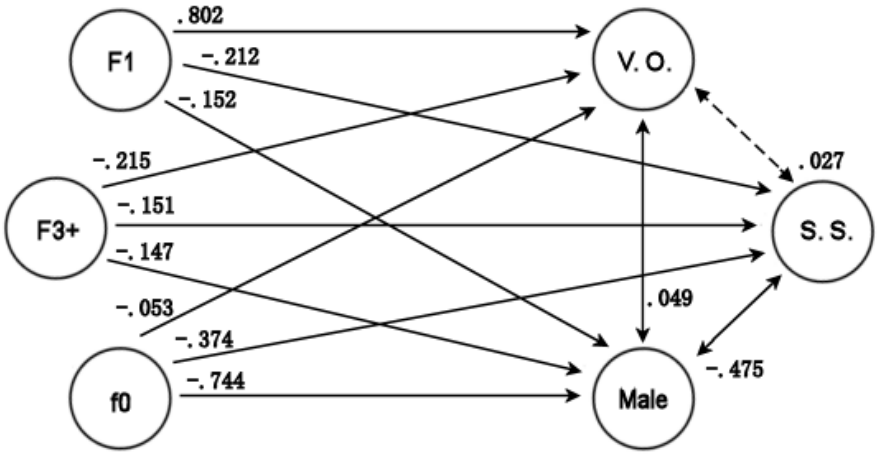

Figure 2.3. *Partial correlation coefficients (averaged over participants) between pairs of variables (V.O. = Vowel Openness, S.S. = Speaker Size, Male = Maleness). The broken line between Size and Vowel Openness indicates the only relationship which did not reach significance by t-test. Arrows indicate the presumed direction of effects.*

Let us define a pure direct relationship between f0 and vowel openness as one that is not mediated by apparent-speaker characteristics. For example, the relationship between f0 and vowel quality in the model of Syrdal and Gopal (1986) qualifies as a pure direct relationship in this sense. The inclusion of concomitant information about a listener's impression of apparent-speaker characteristics should not affect this direct relationship in any way. Specifically, the correlation between vowel quality and f0 would be essentially unaffected after controlling for a listener's judgment of speaker gender in a partial correlation analysis.

Similarly we define a pure indirect relationship between f0 and vowel openness as one that is mediated by the direct effects of f0 on certain apparent-speaker characteristics: f0 affects the apparent-speaker characteristics which in turn affect vowel openness. In such a case, when behavioral measures of those apparent-speaker characteristics are accounted for, the partial correlation between f0 and vowel openness will approach zero.

The cases outlined above represent the endpoints of a range of possibilities. A series of exploratory models were considered that were intended to shed light on the relative direct and indirect effects of the three manipulated variables in the experiment (f0, F1 and F3+) on vowel openness.

To this end, we examined changes in partial correlation coefficients between manipulated variables and vowel openness in two kinds of models. We will illustrate these kinds of model for f0. The first kind of model will be referred to as a fully-controlled model. It is identical in form to the kind of analysis reported in Section 2.3.1. To review, the partial correlation between f0 and vowel openness is calculated after controlling for all other variables; namely the two other manipulated variables F1 and F3+ and the two other response variables Maleness and Speaker Size. The second kind of model will be called the no-speaker model, where the response variables Maleness and Speaker-Size are left out of the model.

Thus the original, fully-controlled model correlations included apparent-speaker characteristics, while the no-speaker model ignores them. Our analysis follows the logic outlined in the beginning of this section. If f0 has a largely direct relation to vowel openness, then there should be little difference in the partial correlations between f0 and openness of the fully-controlled and no-speaker models. If the relation is predominantly indirect, then it is expected that the partial correlation coefficients between f0 and vowel openness will decrease noticeably in magnitude in the fully-controlled model. The degree of this decrease will be taken as a measure of the relative indirectness of the relationship.

Similar assessments of the relative indirectness of the other two manipulated variables, F1 and F3+, were undertaken. For assessing the relation

between F1 and vowel openness, f0 and F3+ were partialled out as control variables; for assessing the relation between F3+ and vowel openness, F1 and f0 were the control variables.

Differences between the two models were tested using the same process outlined in Section 2.3.1, following the two-stage analysis of Lorch and Myers (1990). If the partial correlation coefficients do not change significantly between the two models, the expected value of the differences between the two estimated partial correlation coefficients for a single participant will approach zero. To test this, a series of paired t-tests were carried out on the differences between the two estimated coefficients across the 19 participants. The results of these t-tests show that all three differences are significant indicating that the inclusion of apparent-speaker characteristics in the model significantly affects the relationship between vowel openness and F1, F3+ and f0. Furthermore, in all three cases the partial correlation coefficients as estimated by the fully-controlled model decrease in magnitude relative to those obtained from the no-speaker model indicating that F1, F3+ and f0 all have significant indirect effects on vowel quality. Of the three cues investigated, f0 was most strongly affected by the inclusion of apparent-speaker characteristics (mean difference = 0.091, t = 4.48, df = 18, p-value < 0.0003), followed by F1 (mean difference = 0.021, t = 4.71, df = 18, p-value < 0.0002) and F3+ (mean difference = 0.017, t = 2.573, df = 18, p-value < 0.02).

	F1	**F3+**	**f0**
No-speaker Mean	0.824	-0.232	-0.144
Fully-controlled Mean	0.802	-0.215	-0.052
Decrease in Magnitude	2.6%	7.5%	63.3%

Table 2.4. *Mean partial correlation coefficients across all 19 participants for the fully-controlled and no-speaker models. The percent decrease in mean indicates the decrease in magnitude from the fully-controlled model to the no-speaker model as a function of the magnitude of the no-speaker model.*

Another way to consider changes in the estimated coefficients across the two models is to consider the change in the mean partial correlation coefficient for pairs of variables between the no-speaker and fully controlled models. The means for pairs of variables across both models, and the corresponding percentage decreases in magnitude are presented in Table 2.4. Although the absolute change in the F3+ coefficient is smaller than that seen in the F1 coefficients, when this is considered as a percentage of its original magnitude, the relative change in F3+ is actually larger than that of the F1 coefficients. The change in the f0 coefficients is dramatically larger than either the F1 or F3+ changes. These results reinforce those presented in Section 2.3.1 which suggested that F1 was more strongly related to vowel openness than F3, and that f0 is strongly related to apparent-speaker characteristics but only a weak direct determiner of vowel openness.

2.5 General Discussion

Since this experiment was designed to investigate the relationship between f0 and vowel quality, the first question is whether f0 affects vowel quality at all. It is clear that it does, participants identified an average of 11% more vowels as /ʌ/ when they had the highest f0 relative to the same vowels when presented with the lowest f0. This result is quite far from zero ($t = 6.1254$, $df = 18$, p-value = <.0001) and only 1 of 19 listeners did not show an increase in the number of vowels identified as /ʌ/ as f0 rose. The change in f0 must be ultimately responsible for the change in vowel quality across f0 levels since the vowels across f0 levels are identical in all other respects.

Not only does f0 have an effect on perceived vowel quality, but both sets of partial correlations considered in the previous section show a significant relationship between f0 and vowel quality after adjusting for other factors considered in either model. These results are difficult to reconcile with any hypothesis in which f0 is completely uncorrelated with vowel quality. Smith et al. (2005) and Irino and Patterson (2002) have proposed that vowel quality is entirely determined by aspects of the spectrum independent of f0. Since the partial

correlation between f0 and vowel openness was calculated after correcting for F1 and F3 information[11], and these factors should entirely determine vowel quality, it is not clear why f0 should have such a persistent relationship with vowel quality. In fact, our results indicate that any theory of vowel perception which completely disregards the influence of f0 on vowel quality cannot be an accurate representation of human behavior, at least in these random-speaker listening conditions.

The question then becomes whether the effect of f0 on vowel quality is mainly direct (as is the effect of the FFs) or mainly indirect (as is the effect of apparent-speaker characteristics). If the effect of f0 on vowel quality were direct and based on the natural covariance of FFs and f0s experienced by people on a daily basis, then the relationship between these two variables, all other things being equal, should cluster around the value dictated by this natural covariance; it should not be spread over a large range of values. Additionally, the relationship between f0 and vowel openness should not be dramatically affected by controlling for relevant apparent-speaker characteristics. Specifically, if the relationship of f0 to vowel quality is of the same kind as the relationship between vowel quality and the formants, then the f0-vowel openness relationship and the F1-vowel openness and F3-vowel openness relationships should change in similar ways as a result of controlling for Speaker Size and Gender.

Our results indicate that none of the restrictions or predictions posited by a direct f0 hypothesis play out. Participants show a wide range of sensitivities to this relation, in some cases even showing exactly the opposite relation between f0 and vowel quality that one would expect. Although the behavior of a few participants is unusual or difficult to interpret, the variation exhibited is itself a challenge to any theory of vowel perception in which f0 is tied to vowel quality in a stable and consistent way. If the effect of f0 is not fixed, but is instead modifiable to suit the listening conditions, then it ceases to be direct f0

[11] Apparent-speaker characteristics were also accounted for in the larger model, however, these should not affect the outcome according to f0-free hypotheses.

normalization. This will also apply to any scheme that relies on fixed F1-f0 relations in the determination of vowel quality (see also Johnson 1990). Furthermore, the relationship between f0 and vowel openness is considerably weakened after controlling for apparent-speaker characteristics while the F1-vowel openness and F3-vowel openness relationships maintain much of their strength. Although this does not tell us about the exact relationship between f0 and vowel openness, it is enough to conclude that this relationship is of a different kind than that between the FFs and vowel openness.

The hypothesis that f0 affects vowel quality mainly indirectly, via its effect on apparent-speaker characteristics is perhaps the only remaining viable hypothesis, and its predictions are well-supported by our results. Although f0 strongly affects vowel quality, once apparent-speaker characteristics have been accounted for (using the response variables Speaker Size and Maleness) the relationship between f0 and vowel quality is weakened. Additionally, both Speaker Size and Maleness show a consistent relationship with vowel openness independently of the FFs and f0. It seems that f0 affects vowel quality insofar as it affects a listener's expectations about the presumed speaker. This is so whether such expectations take the form of general characteristics used by traditional normalization theories (e.g., formant ranges or vocal tract length) or the more detailed individual apparent-speaker characteristics of exemplar-oriented models.

However, although the indirect effect of f0 on vowel quality seems to be the more salient one, f0 still appears to exert a significant direct effect on vowel quality. The variables we used to measure apparent-speaker characteristics, Speaker Size and Maleness, were, in effect, surrogates for listener-internal latent variables that specify whatever speaker information directly affects vowel quality. It is possible that the apparently direct effect of f0 on vowel quality might actually be due to the fact that our indices of apparent-speaker characteristics, Speaker Size and Maleness, are not sufficient to fully approximate the true values of the relevant internal variables. However, the results we have presented strongly support a theory of vowel perception in which the presumed identity of the speaker plays an important role in the determination of vowel quality. A more

elaborate form of latent variable modeling and/or a better set of behavioral instruments relating to relevant judgments of apparent-speaker characteristics might elucidate this question.

In the introduction we suggested the normalization process was being approached as a black box system where we would not seek to define the exact internal working of the process but simply to infer what information plays a significant role in the system's transfer characteristics. At this point it seems fair to say that both f0 and apparent-speaker characteristics play a role in this process in a manner broadly consistent with an indirect model of speaker normalization. However, the precise mechanisms by which these factors operate remains to be determined.

Works Cited

Assmann P.F., Dembling S., and Nearey T.M. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. *Proceedings of the Ninth International Conference on Spoken Language Processing*,, Pittsburgh, PA, September 17-21, 2006. 889-892.

Assmann, P. F. and Nearey, T., M.. (2007). Relationship between fundamental and formant frequencies in voice preference. *Journal of the Acoustical Society of America* 122: EL35-EL43.

Assmann, P. F. and Nearey, T., M.. (2008). Identification of frequency-shifted vowels. *Journal of the Acoustical Society of America* 124: 3203- 3212.

Bachorowski, J. and Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America*106: 1054-1063.

Boersma, Paul and Weenink, David (2009). Praat: doing phonetics by computer. Version 5.1.01, retrieved October 2008 from http://www.praat.org/

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behavior* 60, 773–780.

Fant, Gunnar. (1960). Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations. The Hague: Mouton. 107-138.

Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging *Journal of the Acoustical Society of America* 106: 1511-1522.

Fujisaki, H. and Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics AU-16*, 73-77.

Glidden, C. and Assmann, P. F. (2004). Effects of visual gender and frequency shifts on vowel category judgments. *Acoustics Research Letters Online* 5: 132-138.

Gonzalez, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics* 32: 277-287.

Gottfried, T.L. and Chew, S.L. (1986) Intelligibility of vowels sung by a countertenor. *Journal of the Acoustical Society of America 79*, 124-130.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97: 3099-3111.

Hillenbrand, J. M. and Houde, R. A.. (2003). A narrow band pattern-matching model of vowel perception. *Journal of the Acoustical Society of America* 113: 1044-1055.

Hollien, H., Green, R., and Massey, K. (1994). Longitudinal research on adolescent voice change in males. *Journal of the Acoustical Society of America* 96: 2646–2653.

Irino, T. and R. D. Patterson. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. *Speech Communication* 36: 181–203.

Johnson, Keith. (1990). The role of perceived speaker identity in f0 normalization of vowels.

Johnson, Keith, Strand, Elizabeth A. and Mariapaola D'Imperio. (1999). Auditory-visual integration of talker gender in vowel perception. Journal of Phonetics 27: 359-384

Johnson, Keith. (2005). Speaker Normalization in speech perception. In Pisoni, D.B. and Remez, R. (eds) *The Handbook of Speech Perception*. Oxford: Blackwell Publishers. pp. 363-389.

Joos, M. (1948). Acoustic Phonetics. Language 24: 1-136.

Klatt, Dennis H. and Klatt, Laura C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America* 87: 820-857.

Kunzel, H. J. (1989). How well does average fundamental frequency correlates with speaker height and weight?. *Phonetica* 46: 117–125.

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America* 29: 98–104.

Lass, N. J., and Brown, W. S. (1978). Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. *Journal of the Acoustical Society of America* 63: 1218–1220.

Lass, N. J., Phillips, J. K. and Bruchey, C. A. (1980). The effect of filtered speech on speaker height and weight identification. *Journal of Phonetics* 8: 91-100.

Lehiste, I. and Meltzer, D. (1973) Vowel and speaker identification in natural and synthetic speech. *Language and Speech 16*, 356-364.

Lorch, R. F., and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. Journal of Experimental Psychology: Learning, Memory, and Cognition 16: 149-157.

Mendoza, E., N.Valencia, J.Muñoz and H.Trujillo. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). Journal of Voice 10: 59-66.

Miller, J. D.. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America* 85: 2114-2134.

Miller, R.L. (1953) Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America 25*, 114-121.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. . *Journal of the Acoustical Society of America* 85: 2088-2113.

Nearey, T. 1992. Context effects in a double-weak theory of speech perception. *Language and Speech* 35: 153-172.

Nearey T.M. and Assmann P.F. (2007). Probabilistic 'sliding template' models for indirect vowel normalization. Experimental Approaches to Phonology, eds. M. J. Solé, P. S. Beddor, and M. Ohala. Oxford University Press, pp 246-269.

Perry, T. L., Ohde, R., N. and D. N. Ashmead. (2001). The acoustic bases for gender identification from children's voices. *Journal of the Acoustical Society of America* 109: 2988-2998.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 175-184.

Price, P. J. (1989). Male and female voice source characteristics: Inverse filtering results. Speech Communication 8: 261-277.

Rendall, D., Vokey, J. R., and Nemeth, C.. (2007). Lifting the Curtain on the Wizard of Oz: Biased Voice-Based Impressions of Speaker Size. *Journal of Experimental Psychology: Human Perception and Performance* 33: 1208 –1219.

Slawson, A.W. (1968). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America 43,* 87-101.

Smith, David R. R. and Roy D. Patterson. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. Journal of the Acoustical Society of America 118: 3177-3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H. and Toshio Irino. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America* 117: 305-318.

Strand, Elizabeth, A. (2000). Gender Stereotype Effects in Speech Processing. PhD Dissertation, The Ohio State University.

Syrdal, A. K. and Gopal, H. S.. (1986) . A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America* 79: 1086-1100.

Traunmüller, H. (1994). Conventional, biological, and environmental factors in speech communication: A modulation theory. Phonetica 51: 170-183.

Van Borsel, J., J. Janssens, M. De Bodt. (2009). Breathiness as a Feminine Voice Characteristic: A Perceptual Approach. Journal of Voice 23: 291-294.

van Dommelen, W. A. and Moxness, B. H. (1995). Acoustic Parameters in Speaker Height and Weight Identification: Sex-Specific Behavior. *Language and Speech* 38: 267-287.

van Dommelen, W. A. (1993). Speaker height and weight identification: A re-evaluation of some old data. *Journal of Phonetics* 21: 337-341.

Chapter 3

Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis

3.1 Introduction

There is a many-to-many relationship between vowel categories and the acoustic characteristics listeners use to determine vowel quality (Peterson & Barney 1952). Productions of a single vowel category by different speakers might be very different acoustically, just as productions of different vowel categories by different speakers might be very similar acoustically. Differences in production between speakers may arise from differences in speaker gender, size, age, dialect, or any number of other factors. Despite potentially large differences in the acoustic characteristics of a vowel when produced by different people, listeners generally identify vowel tokens with good accuracy. Even for isolated vowels, free from any consonantal context, identification can be quite high (Assmann et al. 1982; Macchi, 1980; Rakerd et al. 1984). However, it is well known that for a given set of listening conditions, speech presented in a mixed-voice condition tends to be identified less accurately and more slowly than when similar stimuli are presented blocked by voice (Assmann et al. 1982; Creelman, 1957; Magnuson & Nusbaum, 2007; Mullennix, Pisoni, & Martin, 1989; Verbrugge et al. 1974; Nusbaum & Morin 1992). The drop-off in identification performance in mixed-voice listening conditions relative to single-voice conditions for the same task will be referred to as the *mixed-voice* effect.

The mixed-voice effect is also associated with additional processing relative to single-voice conditions. Wong et al. (2004) report that listeners

demonstrate increased activity in areas of the brain involved in speech perception in mixed-voice vs. single-voice listening conditions, indicating that mixed-voice listening conditions bear an added cognitive burden. Nusbaum & Morin (1992) asked participants to remember a series of numbers during a speech identification task and found that this increased reaction times only in mixed-voice conditions, indicating that the process of adapting to differences between speakers interacts with working-memory load. Similarly, Martin et al. (1989) found that serial recall of word-lists is worse when the words are produced by multiple voices, relative to when they are produced by a single voice. Although the exact nature of the mixed-voice effect, and the cause of the additional processing observed in mixed-voice listening conditions, is not exactly known, it seems likely to arise from the mechanism by which listeners account for differences between speakers.

The process by which listeners account for speaker-to-speaker differences in the production of vowels is commonly referred to as *normalization*. Many theories of normalization involve the estimation of a speaker-dependent formant-space (Joos, 1948; Ladefoged & Broadbent, 1957; Gerstman, 1968; Ainsworth 1975; Nearey 1978; Nearey, 1989; Nearey & Assmann 2007). The estimate of the speaker's formant-space need only be detailed enough to let the listener know what formant frequencies should be expected for a given vowel category, when produced by the speaker. The listener then determines vowel quality in reference to the estimate of the speaker's formant-space, rather than by considering the acoustic information carried by vowel sounds in absolute terms. Following this tradition, the term *normalization* will be used to refer to the process by which a listener arrives at an estimate of a speaker-dependent formant-space in order to interpret the vowels produced by a speaker.

If normalization were carried out for each vowel token in turn, without reference to what has been heard previously, one would expect that identification rates for vowels produced by a given speaker would not depend on the number of voices in the round. In addition, reaction times associated with the identification of a given set of speech sounds should not vary based on whether they were presented in a mixed- or single-voice listening condition. Instead, the existence of

a mixed-voice effect strongly suggests the importance of extrinsic information in vowel perception, and for the process of normalization[12]. In single-voice blocks, the listener is presented with vowels produced by a single voice so that information from previously heard vowels may be used in order to more accurately identify upcoming vowels. In mixed-voice blocks, the formant-spaces of speakers may differ in such a way that considering vowels produced by one speaker relative to the formant-space of a second speaker may lead to errors. This fundamental difference between mixed- and single-voice listening conditions may help to explain some of the causes of the mixed-voice effect.

3.1.1 Contextual Tuning Theory

Nusbaum & Morin (1992) and Magnuson & Nusbaum (2007) suggest that normalization is controlled by a process they refer to as *contextual tuning*. This approach to normalization is summarized in Nusbaum & Morin (1992):

> "attentional demands increase [in mixed-voice conditions] because the presence of this variability in relationships between speech and linguistic responses requires active processing to reduce the set of possible responses to a single response (Nusbaum & Schwab, 1986). This active processing uses information contained within a single token of speech to provide the context for recognizing the linguistic structure of the utterance, namely a representation of the talker's vocal characteristics. When the listener can develop a mental representation of the talker's vocal characteristics to constrain the representation of subsequent utterances, the demands on attention are reduced." (p. 125).

This formulation of contextual tuning suggests that a listener arrives at a formant-space estimate using information carried by the first speech sound produced by a new voice to interpret subsequent productions by that same voice,

[12] Extrinsic information is information which is not carried by a vowel sound itself, while intrinsic information is carried within the vowel (Ainsworth 1975, Nearey 1989). For example, the average pitch or formant frequencies of a carrier phrase that precedes a vowel is extrinsic to the vowel, while the formant frequencies and pitch of the vowel are intrinsic to it.

and is thus generally compatible with a (conditional) extrinsic-normalization framework. Nusbaum & Magnuson (2007) refine the theory, stating that:

> "a change in talker triggers normalization procedures that operate until a stable mapping between the talker and internal phonetic categories is achieved. The stable mapping is then maintained until a talker change is indicated acoustically (e.g., by large changes in F0) or more implicitly (e.g., via failures of lexical access)" (393).

They later note that:

> "The problem of adjusting to changes in talker characteristics then might be thought of as the same kind of computational problem as recognizing phonetic structure (cf. Nusbaum & Magnuson, 1997). In other words, detecting talker differences that require perceptual accommodation is itself a perceptual problem that may not be handled automatically or passively" (402).

3.1.1.1 An elaboration of the contextual tuning approach

Magnuson and Nusbaum (2007) make it clear that their intent is not to investigate the specific mechanisms involved in normalization or the detection of speaker changes. Their goal is only to investigate the cognitive mechanisms by which the normalization process is controlled, stating, "[t]he heavy lifting of identifying specific mechanisms remains" (406). Although a full-fledged identification of such mechanism will not be attempted here, it is useful to explore some modest extrapolations of this general framework that relate in part to somewhat more specific proposals about normalization from the literature that can be subjected to empirical test.

According to contextual tuning, the important factor governing the use of extrinsic information in vowel perception is not whether there has been an actual speaker change, but whether the listener *thinks* that there has been a speaker change. Because of the many-to-many relationship between the acoustic characteristics of a speech sound and speaker changes, it is difficult to delineate the exact conditions under which a listener will detect a speaker change. For

example, Magnuson & Nusbaum (2007) report an experiment (Experiment 4) in which listeners performed a speeded monitoring task for blocks made up of synthetic voices which differed only slightly in their f0 (150 Hz vs. 160 Hz), but were identical in all other respects. One group of listeners was told that the blocks contained a single voice while the other group was told that the blocks contained multiple voices. The group which was instructed that blocks contained multiple voices exhibited a significant increase in reaction times relative to the group which was told that the blocks contained a single voice. Presumably, listeners who were instructed to expect multiple speakers treated the condition as a mixed-voice one, thereby leading to the longer reaction times typically observed in such tasks. The group which was instructed to expect one voice did not detect speaker changes and did not exhibit the increase in reaction times, despite being presented with identical stimuli.

Contextual tuning is composed of two processes that may result in additional cognitive demands and may help explain the increase in reaction times present in mixed-voice conditions. First, the estimation of the speaker-dependent formant-space may be a cognitively burdensome process, which results in increased reaction times. Although the refinement of the formant-space estimate may be an ongoing process in single-voice conditions, it seems reasonable to think that at some point a listener may become familiar enough with a speaker's voice so that normalization is no longer necessary (i.e., a 'stable mapping' between acoustic input and internal representations has been achieved). In a block in which voices (and their related formant-spaces) change from trial to trial in an unpredictable manner, a listener may never arrive at this level of confidence. Another possibility is that the initial estimation of the formant-space is the most cognitively burdensome, and that refinements to this estimate are less costly. If this were the case an increase in reaction times in mixed-voice conditions would also be observed even if listeners performed formant-space estimations for each vowel since mixed-voice listening conditions would results in relatively more initial estimations than refinements.

Secondly, the detection of speaker changes, or the diversion of some cognitive capacity in order to detect speaker changes, may slow the identification of speech sounds. Although it is reasonable to think that listeners may also monitor for speaker changes in single-voice conditions, this process may not be given a high priority in situations in which listeners do not expect a speaker change. Furthermore, in the event that a speaker change is detected, secondary processes that bear an additional cognitive load may become active. For example, when a likely speaker change is detected, the listener may attempt to estimate the characteristics of the new speaker (e.g., gender, height, age, socioeconomic status, dialect). The listener may also attempt to assess how necessary it is to re-initiate normalization completely, or whether any evidence from previous speech sounds might be used to inform the new re-estimation.

Contextual tuning may also help explain some of the decrease in identification rates for mixed-voice conditions. Because of the uncertainty involved in the detection of speaker changes in a mixed-voice block, listeners may fail to notice a speaker change, just as they might think that there has been a speaker change in cases where there has not. When there are large formant-space differences between speakers, failing to notice a speaker change, and combining extrinsic information from multiple voices, may lead to errors. This suggests that at least some of the decrease in performance associated with the mixed-voice effect is due to the inability of listeners to correctly detect speaker changes in situations where it would be beneficial to do so to maintain high identification accuracy. If this view of normalization is correct, then one would expect that in situations that facilitate the detection of speaker changes, the decrease in accuracy related to formant-space differences between speakers might be minimized.

Although not explicitly stated by Nusbaum and colleagues, contextual tuning seems to imply a rather complex relationship between reaction times, identification accuracy and the detection of speaker changes. In general, phonetically ambiguous stimuli, or more difficult mixed-voice lists, might be expected to result in a decrease in accuracy and an increase in reaction times so that identification accuracy and average reaction times may be negatively

correlated across blocks (see Whalen et al. 1993). Independently of this relationship, the detection of speaker changes and the re-initiation of the normalization process may also result in an increase in reaction times. However, since the re-initiation of the normalization process resulting from a detected speaker change should result in a more accurate estimation of the speaker's vowel space, it should result in relatively higher identification accuracy by reducing ambiguity. Consequently, if contextual tuning is correct, one would expect that when the listener detects a speaker change, reaction times will increase without necessarily being associated with lower accuracy.

3.1.1.2 Differential predictions of alternative accounts

This version of contextual tuning may be contrasted with two alternative views of normalization in which the detection of speaker changes does not play an important role. In *pure-intrinsic* normalization theories, the detection of speaker changes is irrelevant because extrinsic information does not play an important role in vowel perception (Syrdal & Gopal, 1986; Smith et al. 2005). According to these views, each vowel token is essentially 'self-normalizing' in that it carries all the information necessary for its interpretation. If this were the case, we would expect that identification rates for vowels for a given voice should not vary based on whether they were presented in a mixed- or single-voice condition. With respect to reaction times, although listeners may take more or less time to identify a given vowel produced by a certain voice, there is no clear reason why the reaction times associated with the identification of a set of stimuli should vary systematically based on whether they are presented in a mixed- or single-voice condition. Furthermore, although there may be a positive relationship between average reaction times and identification accuracy in a block, this relationship should not be mediated in any way by the detection of speaker changes.

A second possibility is that extrinsic information is important, but that listeners use information related to the spectral properties of the last *n* tokens (or the last *k* seconds of speech) in order to estimate the speaker-dependent formant-space, with no role for the detection of speaker changes. This might be expected if normalization were primarily driven by mechanisms such as those reported in

Watkins & Makin (1994) and Watkins & Makin (1996), in which listeners were demonstrated to compensate for the long-term spectral characteristics of a signal when identifying vowel sounds. In a series of experiments, Watkins and Makin presented listeners with a carrier phrase followed by a word containing a vowel token, and asked listeners to identify the word that followed the carrier phrase. Several experiments were carried out, and several different filters were applied to the carrier phrases.

Results indicate that the perceived identity of the vowel following that carrier phrase was predictable based on the long-term average spectral characteristics of the carrier phrase. The authors suggested that some of the perceptual shifts observed in experiments which manipulate carrier phrases to affect the perceived identity of a following target may be caused by accommodation to the long-term average spectral characteristics of the carrier phrase, and not the result of the listener adapting to the formant-space suggested by the carrier phrase. Although there are no clear examples of a normalization method that relies solely on a mechanism like this in the literature, a formant-space normalization system that utilizes statistics such as formant means or ranges over given intervals might have generally similar properties.

A normalization method which worked solely by mechanisms of this kind might be termed *passive-extrinsic*, since the extrinsic information involved in the process is not variable based on perceived speaker changes or listener expectations, but only on the recent history of stimulus properties (in contrast to this, contextual tuning might be thought of as an *active-extrinsic*[13] model of normalization). If the estimate of the speaker-dependent formant-space involved the joint consideration of information from a fixed number of previous tokens, identification errors would be correlated with the difference between the formant-spaces of the two voices, since the estimated formant-space would be somewhere

[13] The distinction between active and passive control structures, and their implications for theories of normalization is discussed in detail in Magnuson & Nusbaum (2007). In short, active control structures allow for the same input to result in different outputs based on the specific listening situation, while passive control structures feature a predictable and rigid relationship between input and output regardless of context.

between these two. Reaction times might be expected to vary based on the phonetic ambiguity of the vowels being presented, but again, there should not be systematic variation in the relationship between reaction times and identification accuracy resulting from whether the listener thought the round contained one, or more than one speaker.

3.1.1.3 Testing Contextual Tuning Theory in Magnuson & Nusbaum (2007)

Magnuson & Nusbaum (2007) present the results of an experiment (Experiment 1) meant to offer explicit support for contextual tuning theory[14]. The stimuli consisted of isolated vowels produced by four natural voices; those of two adult males and two adult females. The average F1 and F2 values for the vowels of the two female speakers differed by only 0.3%, while the average F1 and F2 values for the two male speakers differed by 5.4%. Although within-gender differences were somewhat larger for the males than for the females, both were small compared to the 20% differences between males and female speakers.

Vowels were presented in blocks of 16 vowels produced by either a single voice, or two different voices. Each listener heard vowels presented in both single- and mixed-voice conditions, where one group of listeners was always presented with mixed-voice blocks in which speakers were of the same gender, and another group was presented with mixed-voice blocks in which speakers were of different genders. Within each block, the target vowel was one of /i ɪ u ʊ/, while distractors were chosen from the vowels /e æ ʌ ɛ/, plus any of the four target vowels that were not acting as targets for that particular block. Each block contained a total of four targets inserted randomly into the sequence, with the constraint that no two targets appear in a row. Listeners performed a speeded-monitoring task where they had to push a computer key as soon as they heard the

[14] This experiment is a replication of Experiment 4 in Nusbaum & Magnuson (1992). The pattern of results reported for that experiment are generally consistent with what is reported in Experiment 1 of Magnuson & Nusbaum (2007). Unfortunately, the authors do not provide a full accounting of results, nor do they provide a useful description of their vowel stimuli. For those reasons, the results of that experiment will not be discussed here.

target vowel (indicated to them on a monitor), and ignore all non-target distractor vowels. Response times were measured from stimulus onset, and hit rates (responses registered following targets) and false alarms (responses registered following distractors) were collected.

Magnuson & Nusbaum report a significant decrease in hit rates for mixed-voice blocks relative to single-voice conditions. Hit rates were slightly higher for different-gender blocks relative to same-gender blocks overall, but the main effect for gender homogeneity did not reach significance. There was a nearly significant ($p = .072$) interaction between talker condition (mixed-speaker vs. single-speaker) and gender homogeneity. Reaction times were significantly higher in all mixed-voice blocks relative to the single-voice blocks, save for the female-female mixed-voice blocks which did not differ significantly from single-voice blocks.

According to contextual tuning, performance may be higher in different-gender mixed-voice blocks than in the same-gender mixed-voice block because listeners are aware that these blocks contain multiple speakers. This realization may partly counteract the negative effect of the much larger formant-space differences between speakers of different genders compared to speakers of the same gender.

On the other hand, although there were relatively smaller differences between the formant-spaces of different speakers of the same gender, listeners may not have realized that the blocks involved multiple speakers; or, even if they did, they may have missed exactly when speaker changes were occurring. As a result, the same-gender mixed-speaker blocks manifested a trend toward slightly lower performance than the different-gender mixed-voice blocks. This is true despite the fact that formant-space differences between voices are smaller in same-gender cases. Finally, although the female-female mixed-voice blocks objectively consisted of vowels from two different voices, reaction times did not differ significantly from those of single-voice blocks, suggesting that listeners may not have realized that the blocks contained more than one speaker. This highlights the fact the detection of speaker changes is an imperfect, non-deterministic process.

Although the trends in the pattern of results are generally consistent with contextual tuning theory, many effects tested in Magnuson & Nusbaum (2007) are generally weak or non-significant and thus do not offer strong support for contextual tuning. However, some aspects of the experimental design may have contributed to the limited size of the effects. First of all, the target vowels used (/ i ɪ u ʊ /) may not be very confusable with each other in mixed-voice conditions. These four vowels were identified correctly in 97% of cases in data presented by Peterson & Barney (1952) and in 98% of cases in Hillenbrand et al. (1995). Furthermore, the vowels which are most spectrally similar / u ʊ / and / i ɪ / may be distinguishable on the basis of durational differences or because of vowel inherent spectral change when produced by natural voices (Hillenbrand et al. 1995; Nearey & Assmann, 1986). Perhaps as a result of this hit rates hovered around 93% in all listening conditions. This leaves very little room to model variation in performance as a result of different voice pairs. Furthermore, because natural voices were used, it is difficult to know which aspect of the speakers' voices listeners were using to detect speaker changes, or under what conditions they were likely to detect speaker changes.

3.1.2 Rationale for current experiment

The experiment to be described below adopts the same basic design used in Experiment 1 of Magnuson & Nusbaum (2007) with some modifications which may enhance and clarify the effects reported for that experiment. A series of synthetic voices was created which differed in their formant-spaces and/or their source characteristics, and the four vowels / æ ʌ ʊ ɑ / were synthesized for each voice. As opposed to the vowels used in Magnuson & Nusbaum (2007), these vowels are generally more difficult to identify: in data presented by Peterson & Barney (1952) they were identified correctly in 93% of cases, while they were identified correctly in 93.7% of cases in Hillenbrand et al. (1995). This was expected to result in lower performance overall. Synthetic voices were used in order to control for random variation in the production of vowels and to eliminate idiosyncratic differences in source characteristics between voices. Furthermore,

each block contained a higher number of total vowel tokens (30) and target tokens (12), in order to allow for more variation in hit rates.

Differences in source characteristics between voices in a block were expressly intended to facilitate the detection of speaker changes in a block, thereby potentially mitigating the decrease in hit rates associated with mixed-voice listening conditions by strongly encouraging the detection of speaker changes when the voices had different sources. The formant-space differences between speakers were intended to result in decreased performance (i.e., the mixed-voice effect) when listeners were unlikely to detect speaker changes in a block (e.g., in the absence of source differences between voices). If a version of contextual tuning theory adequately describes the process of normalization, three general results are expected:

A) The decrease in identification rates associated with formant-space differences in mixed-voice conditions will be mitigated in situations in which the detection of speaker changes is facilitated.
B) In situations where speaker changes are not detected, performance should improve in blocks where voices have similar formant-spaces. When listeners are likely to detect speaker changes (e.g. in blocks with heterogeneous sources), their ability to refine their speaker-dependent formant-space estimate may be limited. This may result in a lack of improvement throughout a block or in lower performance overall.
C) Although average reaction times may co-vary negatively with hit rates for blocks, blocks in which speaker changes are likely to be detected may exhibit an increase in average reaction times without a concomitant decreases in hit rates.

3.2 Methodology

3.2.1 Participants

Participants were 71 native speakers of Canadian English, drawn from the linguistics participant pool at the University of Alberta. Participants received

partial course credit for taking part in the experiment. Participants were randomly assigned to a target vowel group and each participant only monitored for a single vowel. There were 18 participants in each of the target vowel groups, except for the / æ / group which had only 17 participants.

3.2.2 Stimuli

The vowels used in the experiment were / æ ʌ ʊ ɑ /, where one of the four acted as the target and the others acted as distractors. A series of 6 synthetic voices were created which differed in terms of their vowels spaces and/or f0 and source characteristics. Formant-space differences were manipulated by using three formant frequency (FF) scaling levels: a baseline level with FFs appropriate for an adult male, a second level with a 10% increase to all FFs (F1 – F10) and a third level with a 20% increase to all FFs (F1 – F10) relative to baseline. The baseline FF values used are presented in Table 3.1, and these were based on production values collected from native-speakers of Edmonton English. Baseline F4 values were set at 3500 for all vowels with subsequent FFs set to 1050 Hz greater than the previous FF. Formants above F3 were scaled by the same factor as F1 to F3 for the other conditions.

	Baseline FF Values (in Hz)			
Vowel	**æ**	**ʌ**	**ʊ**	**ɑ**
F1	717	665	483	651
F2	1497	1283	1093	1055
F3	2319	2318	2272	2323

Table 3.1. *Formant frequencies for the vowels of the baseline voice.*

The two voice source levels consisted of an f0 of 120 Hz with modal source characteristics and an f0 of 240 Hz with breathy source characteristics. The breathy source characteristics were simulated by setting the source bandwidth to 75 Hz and using 10 dB of negative spectral tilt at 3000 Hz (Klatt & Klatt, 1990). Since f0 level and source characteristics were perfectly correlated, the different f0

and source levels will simply be referred to as voice source characteristics. All vowels had steady-state formants, were 200 ms in duration and were synthesised at a sampling rate of 22050 Hz.

3.2.3 Procedure

The general design of the task is an extension of experiments outlined in Nusbaum & Morin (1992) and Magnuson & Nusbaum (2007). Listeners were asked to perform a speeded monitoring task where they had to respond only when they heard a specific target vowel and ignore all distractor vowels. Each listener monitored for a single target vowel so that the designation of a vowel as either target or distractor is listener-specific. All listeners were told which vowel they would be targeting and which vowels would serve as distractors.

Listeners were presented with all combinations of voice pairs, presented in blocks. There were 21 unique voice pair combinations and listeners heard each combination twice resulting in 42 blocks per participant. Listeners were told that any given block might contain vowels from a single voice or from more than one voice. Thirty vowels were presented within each block, consisting of 6 targets and 9 distractors from each voice (3 instances of each non-target vowel). Vowels were randomized within a block subject to the constraint that no two targets appear in a row. The onset of each vowel within a block occurred one second after the onset of the previous vowel, meaning that each block of vowels was roughly 30 seconds in duration. When a block was completed, there was a self-timed pause, which ended when the participant pressed a button. Reaction times (measured from stimulus onset) and accuracy for responses to targets (hits) and distractors (false alarms) were recorded within a block. The hit rate for a block was calculated by dividing the number of correct identification of targets by the total number of targets in the block. False alarm rates were calculated by dividing the number of responses to non-target distractor vowels by the total number of non-target distractors in the block. The experiment was carried out using DMDX (Forster & Forster 2003), and responses were collected using a USB gamepad.

Although the relatively large source differences between voices were intended to strongly suggest to listeners that there were multiple voices in a block,

while conducting the experiment it was realized that it would be beneficial to ask participants how many voices they thought they heard in a given round. The last 14 participants performed an additional task where at the end of each block they were asked to report whether they thought the block contained one or more than one voice and whether they were confident or uncertain of the number of voices in the block. Participants were told that they would be asked to perform this task at the end of each block prior to the beginning of the experiment. The results from this secondary task strongly met expectations regarding the expected relationship between source differences between voices and the detection of speaker changes. The results of this secondary task, in addition to a summary of tests of heterogeneity of results between participants who completed the secondary task and those who did not, are presented in Appendix 2.

3.3 Results

Since the task was designed to be difficult, participants were screened to ensure that they were completing the task to a minimally-satisfactory level. This was done by removing any participant who had more false alarms than correct identifications of targets. This resulted in the removal of six of 71 participants, 5 from the /ʌ/ target group, and one from the / ɑ / target group. All further discussion will be based on the results of the remaining 65 participants.

Each participant heard a total of 1,260 vowels across all 42 blocks for a total of 81,900 trials across all participants. Since the software used only registered one response per stimulus, very fast responses were ambiguous. For example, in some cases responses were registered only 10 ms after stimulus onset, making it more likely that it was a very late response to the previous stimulus than a very fast response to the current one. As a result of this, when a reaction time under 200 ms (the duration of the vowel stimuli) was registered, both the stimulus that was responded to and the stimulus that immediately preceded it were discarded. Participants responded in less than 200 ms in 533 cases, resulting in 1,065 discarded responses (1.3% of total responses) and 80,835 useable trials. An

average of 16.4 responses were lost from each participant (SD = 14.2) with the most lost from any participant being 64 trials, 5% of total trials for that participant.

The predictions made by the contextual tuning hypothesis (outlined at the end of Section 3.1.2) relate directly to the formant-space and source differences between voices in a block. To test these predictions more directly, all blocks were classified into one of six voice-pair types based on the acoustic differences between the voices in the block --i.e., formant-space differences of 0%, 10% or 20% between voices, and either homogeneous or heterogeneous voice sources for each formant-space difference. Hit rates, false alarm rates and average reaction times (for correct identifications) were calculated for each block, independently for each listener. The average of each of these values was then found for each voice-pair type for each participant, resulting in 18 measurements per listener: an average hit rate, an average false alarm rate, and an average reaction time for each of the six voice-pair types. Unless otherwise specified, the remaining discussion will involve average performance, within-participant, between voice-pair types. Each of the predictions to be tested will be dealt with in turn in the following three subsections (3.3.1 through 3.3.3).

3.3.1 Vowel Identification Performance

A series of repeated-measures analyses of variance was conducted on hit rates, false alarm rates and average reaction times for the two factors used to differentiate voice-pair types: formant-space difference between voices (0%, 10%, 20%) and voice source homogeneity. The average within-participant hit rate, averaged across all voice-pair types, was 76% (sd = 15%) with a minimum of 34% and a maximum of 95% across participants. The distribution of hit rates, organized by voice-pair type, is presented in Figure 3.1. The main effects for voice source homogeneity [$F(1,64) = 4.74$, $p = 0.0331$], and formant-space difference [$F(2,128) = 70.83$, $p < 0.0001$] were both significant, as was the interaction of the two [$F(2,128) = 31.83$, $p < 0.0001$].

The nature of the interaction effect was explored by simple main-effects analysis of hit-rates. When voices in a block had homogenous source characteristics, there was a very strong effect for formant-space differences in hit

rates [F(2,128) = 87.22 p < 0.0001]. As seen in Figure 3.1, the interaction pattern suggests that formant-space differences between voices appear to affect hit-rates less for heterogeneous-source blocks. Despite this reduction, the simple main effect of formant-space differences for heterogeneous source blocks is still significant [F(2,128) = 10.59, p < 0.0001].

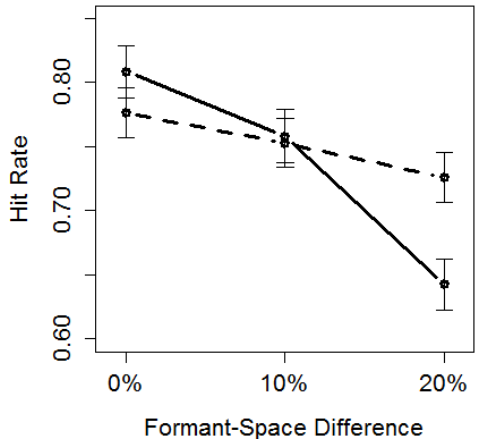

Figure 3.1. *Average within-participant hit rate, presented by voice-pair type. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean.*

Consider now the simple main effects of source-heterogeneity within levels of formant-space difference. Voice source heterogeneity between voices in a block is associated with a 3.2% decrease in hit rates for the 0% formant-space difference [t(64) = 3.33, p = 0.0014], however, when the formant-spaces of voices differ by 10%, source differences between voices have no significant effect on hit rates [t(64) = 0.54, p = 0.56]. When the formant-space of voices differ by 20%, hit rates are 8.3% higher in cases where source characteristics are heterogeneous [t(64) = 5.9, p < 0.0001]. Note that in this case, the effects of source heterogeneity are in the opposite direction from those in the 0% formant-space case, resulting in the crossing lines in Figure 3.1.

Turning now to false alarms, the average within-participant rate was 8.2% (sd = 7.3%) with a minimum of 0.1% and a maximum of 27.3% across participants. A significant main effect for both voice source [F(1,64) = 16.85, p = 0.0001] and formant-space differences was found [F(2,128) = 5.76, p = 0.004].

Unlike the analysis of hit rates, however, the interaction between the two did not reach significance [$F(2,128) = 1.94$, $p = 0.1473$]. On average, source differences between voices in a block resulted in 1.6% more false alarms [$t(64) = 4.1$, $p = 0.0001$]. Formant-space differences of 10% did not significantly increase the number of false alarms relative to blocks in which voices had the same formant-space [$t(64) = 1.1$, $p = 0.26$]; but, when formant-spaces differed by 20% false alarms increased by 1.2% [$t(64) = 3.14$, $p = 0.0025$].

A pattern similar to the false-alarms results was found for reaction times. There was a significant main effect for voice source homogeneity [$F(1,64) = 75.43$, $p < 0.0001$] and formant-space difference [$F(2,128) = 10.59$, $p < 0.0001$], but there was not even a hint of a significant interaction between the two [$F(2,128) = 1.4$, $p = 0.2496$]. The average, within-participant reaction time for the voice-pair type in which voices had the same formant-space and source characteristics was 516 ms (sd = 62 ms), with voice source heterogeneity resulting in an average delay of 27 ms [$t(64) = 8.7$, $p < 0.0001$]. Compared against the control 0% formant difference case, formant-space differences of 10% resulted in an added delay of 10.9 ms [$t(64) = 4.1$, $p = 0.0001$] relative to blocks with no formant-space differences, while formant-space differences of 20% resulted in an added delay of 12.4 ms [$t(64) = 4.3$, $p < 0.0001$] relative to blocks with no formant-space differences. There was no significant difference in response times between blocks with 10% and blocks with 20% formant-space differences [$t(64) = 0.46$, $p = 0.64$].

3.3.2 Improvement within a block

According to contextual tuning (at least as elaborated in Section 3.1.1), in blocks where listeners do not detect speaker changes, they are expected to refine their estimate of the speaker-dependent formant-space throughout the block. When voices in a block share a formant-space, this should lead to an improvement in performance from the beginning to the end of the block, as every consecutive token provides the listener with information which may be used to accurately refine their estimate. On the other hand, in cases where the listener is likely to detect speaker changes, they are expected to re-initialize the normalization process and avoid the use of inappropriate extrinsic information in normalization.

This is expected to mitigate some of the mixed-voice effect, by minimizing the inappropriate use of extrinsic information. However, it may also mean that listeners are not able to refine their estimate of the speaker-dependent formant-space as the block progresses to the extent that they would in the absence of detected speaker changes.

An analysis was devised to summarize the nature of change of identification accuracy during the course of a block and to relate patterns of such change to voice-pair type. Each block contained a total of 30 vowels, 12 of which were targets. Although the targets within a block were presented in a random order (with the constraint that no two targets appear in succession), targets can be considered in terms of the order in which they appeared in a block. On average, in cases where the performance of listeners improves in a block, hit rates for target n_i is expected to be lower than performance for target n_{i+1}. In cases where performance decreases throughout a block, performance for target n_i is expected to be higher than performance for target n_{i+1}. When the performance of a listener is stable within a block, there should be no relationship between target position within a block and expected performance for that target. As a result, the slope coefficient relating hit rates to within-block target number should give an indication of how performance varies within a block, with a positive coefficient indicating improvement, a negative coefficient indicating worsening performance and a coefficient of zero indicating stability.

To investigate how performance within a block varies by voice-pair type, all blocks were sorted by voice-pair type, according the acoustic differences between the voices in the block. Targets were assigned a number from 0 to 11, based on the relative position in which they appeared within the block. This target number was then divided by eleven so that target numbers corresponded to equal fractional increments from 0 to 1. In this way, estimated coefficients have a straightforward interpretation as the expected increase in hit rates (measured in percentage points) from the first target in the block to the last target in the block. Within-participant hit rates were calculated, for each target position within each voice-pair type. A regression was then carried out independently for each voice-

pair type and individually for each participant, predicting hit rates by relative target position. This resulted in six estimated coefficients for each participant (one for each voice-pair type). The distribution of these coefficients, organized by voice-pair type, is presented in Figure 3.2.

A repeated-measures analysis of variance was carried out on these estimated coefficients, with voice source homogeneity and formant-space differences (0%, 10%, 20%) between voices in a block acting as within-subjects factors. A significant main effect was found for formant-space differences between voices [$F(2,128) = 6.67$, $p = 0.0017$]. The main effect for voice source [$F(1,64) = 0.81$, $p = 0.3718$] was not significant. Although the interaction between formant-space differences and voice source [$F(2,128) = 2.93$, $p = 0.0573$] fell just short of the conventional .05 significance level, it seemed reasonable to investigate it further. Accordingly, simple main effects tests were performed.

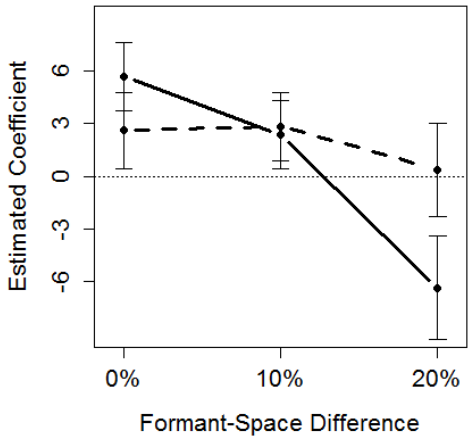

Figure 3.2. *Average coefficient relating within-block target number, and expected hit rates for that target within a block. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean.*

Consider the simple main effect of formant space within source condition. When source characteristics in a block were homogenous, there was a strong effect for formant-space differences on improvement in a block [$F(2,128) = 8.41$, $p = 0.0004$]. However, when voices in a block had heterogeneous source characteristics, there was no significant effect for formant-space differences on improvement [$F(2,128) = 0.49$ $p = 0.6148$]. In these cases, coefficients did not

differ significantly from zero in any case, regardless of formant-space differences between voices, although in all three cases they were slightly positive.

Consider now the case of homogeneous source characteristics. In cases where voices had the same source characteristics and formant-spaces, listeners showed a significant improvement as blocks progressed [$m = 5.7$, $t(64) = 2.96$, $p = 0.0043$], while in cases where voices in a block had the same source but formant-spaces differed by 20%, listeners performed significantly worse as blocks progressed [$m = -6.3$, $t(64) = -2.16$, $p = 0.0345$]. When voices had homogenous source characteristics and a 10% formant-space difference, there was no significant change in hit rates as the block progressed [$m = 2.4$, $t(64) = 1.23$, $p = 0.22$].

3.3.3 The relationship between reaction times, hit rates and the detection of speaker changes

As mentioned in the introduction, phonetically-ambiguous stimuli may take longer to identify in general than less ambiguous stimuli. Since ambiguous vowels should be less accurately perceived, this should by itself result in a negative relationship between the average reaction times in a block and the hit rate for that block. There is in fact a negative relationship between the hit rates and average reaction time in a block. Correlation coefficients between these two measures were calculated for each participant. A between-participants t-test conducted on these correlation coefficients reveals a highly significant negative correlation, averaging -0.18 [$t(64) = -8.5$, $p < 0.0001$].

However, contextual tuning posits that when a speaker change is detected, processes related to the more accurate identification of vowels (e.g., the re-initiation of normalization) are also expected to result in an increase in reaction times. As a result, in situations where listeners are likely to detect speaker changes in a block, reactions times should be higher overall without necessarily being associated with a decrease in hit rates.

To explore how the relationship between acoustic differences and reaction times may be mediated by the detection of speaker changes, the following procedure was carried out individually for each participant. The average reaction

time for each block was regressed on the hit rate for that block, resulting in a reaction-time residual for each block. This residual represents variation in average reaction times that cannot be accounted for by the hit rate for that block. A positive residual indicates that a listener responded slower than expected given their average accuracy, while a negative residual indicates that listeners tended to respond faster than expected given their average accuracy. The mean reaction time residual for each voice-pair type was found, resulting in six measurements for each of the 65 participants. The distribution of average within-participant residuals, grouped by voice-pair type, are presented in Figure 3.3.

Since heterogeneous source characteristics between voices in a block are strongly associated with the detection of speaker changes, it is expected that average reaction times for blocks in which voices have heterogeneous source characteristics should be longer than expected given the hit rate for the block. This suggests that if contextual tuning is correct, the average residual resulting from the analysis presented above should be positive when there are source differences in a block, indicating delays not explicable by ambiguity as indexed by decreased hit rates. The results presented in Figure 3.3 confirm this expectation.

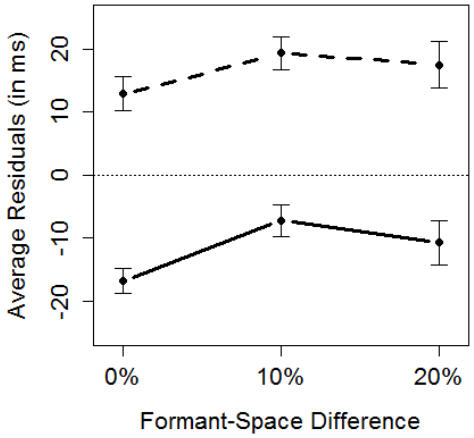

Figure 3.3. *Average, within-participant residual resulting from regressing reaction time on hit rates, presented by voice-pair type. Blocks with homogeneous source characteristics are indicate by a solid line, blocks with heterogeneous source characteristics are indicated by the broken line. Error bars indicate standard errors for each mean.*

To test for the significance of this effect, a two-way, repeated measures analysis of variance was carried out on the average reaction-time residual, with

voice source homogeneity and formant-space difference between voices (0%, 10%, 20%) acting as within-participant factors. A significant main effect was found for voice source [$F(1,64) = 88.02$, $p < 0.0001$], with the average absolute difference in residuals resulting from voice source heterogeneity being 28 ms. The main effect for formant-space differences [$F(2,128) = 3.5$ $p = 0.0331$] was also significant, however, the interaction between voice source and formant-space difference was not significant [$F(2,128) = 0.15$ $p = 0.8588$]. Although formant-space differences affect the reaction time residuals (likely reflecting the fact that these alone were sometimes sufficient to trigger the detection of speaker changes), on average, listeners respond faster than expected given their hit rates when there is voice source homogeneity in a block.

3.4 Discussion

In the introduction, contextual tuning theory was outlined and contrasted with two alternate views of normalization. Rather than focus on the specific processes involved in normalization, these theories were framed in terms of how normalization is controlled, and specifically, how extrinsic information is used in normalization. The two types of theories considered in alternative to contextual tuning theory were pure-intrinsic theories, in which extrinsic information plays no important role in normalization, and passive-extrinsic theories, in which extrinsic information is used by the normalization process in a rigid way. Although they differ in terms of the role played by extrinsic information, both of these alternatives are cognitively passive, in that they do not necessarily require active cognitive control to be carried out (Magnuson & Nusbaum 2007). Furthermore, neither of these alternatives involves the detection of speaker changes in any way. Thus, they cannot predict any relationship between hit rates, reaction times and the detection of speaker changes.

In contrast, contextual tuning theory posits that the detection of speaker changes plays a critical role in guiding listeners' use of extrinsic information in normalization. In a sense, contextual tuning might be thought of as consisting of two 'modes', one being more similar to pure-intrinsic normalization and the other being more similar to passive-extrinsic normalization. In the absence of a detected

speaker change, the listener is in a passive-extrinsic normalization mode and extrinsic information from previous tokens is accumulated and used to identify subsequent vowel tokens. If the formant-spaces of the voices in the block are the same or similar, this refinement will facilitate identification. If the voices in a block have substantially different formant-spaces, the joint consideration of information from different voices may negatively affect hit rates. On the other hand, when a speaker change is detected, the listener shifts to a strategy similar to pure-intrinsic normalization. Previous extrinsic evidence may be discarded as inappropriate and the hit rates associated with a particular vowel token may be closer to those that would be predicted based on the intrinsic properties of the vowel sound.

The experiment described above relied on source differences between voices in a block to give listeners the impression that a block contained multiple voices. The results presented in Appendix 2 confirm this expectation; when voices in a block had homogenous source characteristics listeners were very likely to hear a single voice in a block. As a result, when voices in a block had homogenous source characteristics, listeners may have been in a passive-extrinsic normalization mode. This resulted in good performance when voices in a block had the same formant-space, and poor performance when voices in a block had very different formant-spaces (these two situations are presented in the extreme points on the solid line in Figure 3.1).

In addition, when voices in a block had homogenous source characteristics and the same formant-space, hit rates improves significantly within a block. This suggests that listeners were, in fact, refining their formant-space estimates on the basis of additional extrinsic information in order to arrive at more accurate estimates. On the other hand, when the formant spaces of voices differed by 20%, hit rates declined significantly within blocks, suggesting that identification accuracy suffered from the incorrect combination of extrinsic information from multiple voices.

The variation in hit rates within a block may be explained by the amount of extrinsic information available to a listener for each consecutive vowel target in

a block. For example, the average ordinal position of the first target in a block was 1.6 (out of 30), while the average ordinal position of the final target in a block was 29.1. Clearly, in blocks where voices have different formant-spaces, the chances that a target has been preceded by inappropriate extrinsic information is fairly low for the first target in a block, while it is a certainty for the final target in the block. As a result, in situations in which listeners were unlikely to detect speaker changes, the incorrect use of extrinsic information may increase or become more likely as a block progresses, and identification accuracy may suffer. Conversely, in situations in which voices had the same formant-spaces, listeners would have been provided with increasing amounts of appropriate extrinsic information as a block progressed and the lack of detected speaker changes worked in their favor.

In blocks in which voices had heterogeneous source characteristics, listeners overwhelmingly reported hearing multiple voices in a block. This greatly diminished the negative effect of formant space differences between voices in a block, as demonstrated by the relative lack of change in hit rates when voices in a block had heterogeneous source characteristics (represented by the broken line in Figure 3.1). As opposed to blocks where voices had homogenous source characteristics, hit rates where relatively stable, with no significant increase or decrease in hit rates within a block regardless of the formant-space differences between voices. These results support the notion that, in the presence of detected speaker changes, listeners were likely to be operating in something more similar to a pure-intrinsic normalization mode in which extrinsic information plays a diminished role.

Contextual tuning also suggests a complicated relationship between reaction times, hit rates and the detection of speaker changes. The results presented in Section 3.3.3 indicate that although reaction times are negatively correlated with hit rates, blocks in which voices had heterogeneous source characteristics tended to feature slower reaction times without being associated with decreased hit rates. When considered together with the fact that source heterogeneity resulted in the detection of multiple speakers, a decreased negative

effect of formant-space differences between voices, and stability in identification rates within blocks, this is considered to be strong support of the claim that the detection of speaker changes results in additional processing associated with the more accurate perception of speech, and the avoidance of the incorrect use of previously heard extrinsic information.

Magnuson & Nusbaum (2007) report an increase in reaction times of 29 ms in mixed-voice blocks relative to single-voice blocks for a task very similar to the one reported here (Experiment 1). This is very close to the 27 ms average increase in reaction times resulting from source differences between voices in a block, presented in Section 3.3.1. This suggests that source differences between synthetic voices used here resulted in remarkably similar processing costs to those incurred when listeners are presented with mixed-voice lists consisting of vowels produced by different human speakers in a similar task. Furthermore, this increase in average response times is very close in magnitude to the 28 ms difference in average residuals after controlling for hit-rate resulting from voice source heterogeneity between voices in a block, reported in Section 3.3.3. Since these residuals represent variation in reaction times that cannot be accounted for by the phonetic ambiguity of tokens in a block, this suggests that increases in reaction times resulting from source differences between voices in a block may primarily result from additional processing associated with the detection of speaker changes.

3.5 Conclusion

Taken together, the results outlined in the previous section offer strong evidence for a version of contextual tuning theory as the mechanism that controls the normalization process. Source differences between voices in a block resulted in the impression that there were multiple voices in a block. These differences also resulted in increased reaction times that cannot be fully explained by increased phonetic ambiguity (as indexed by lower hit rates). This is consistent with the hypothesis that the additional processing in blocks with heterogeneous voice sources is actually related to the more accurate perception of many of the vowels. For homogeneous source blocks, the absence of the additional processing associated with a detected speaker change resulted in good accuracy (with

improvement within a block) when voices had similar formant-spaces, and poor accuracy (with decline within a block) when voices had dissimilar vowel spaces. In heterogeneous source blocks, when the listener was more likely to be aware of speaker changes in a block, performance was relatively stable within a block and the negative effect of formant-space incongruences between voices was greatly reduced.

To sum up, the complex pattern of results for hit-rates and reaction time differences outlined above cannot be explained either: a) by a pure-intrinsic normalization process where extrinsic information plays no role whatsoever or b) by an extrinsic normalization in which information is used in a rigid, automatic, fully stimulus-driven manner. By contrast, all the results are reasonably explained by the contextual tuning hypothesis as elaborated in Section 3.1.1.1 and in the discussion. This is a version of contextual tuning that includes a switch between two processing modes guided by the presence or absence of the detection of a change in speaker. The first mode is operative when a new trial is detected as originating from a speaker that is different from that of an immediately preceding trial. It is viewed here as a form intrinsic normalization, where the current speaker's formant-space is estimated only from information in the current utterance and where that fresh estimate is used the identification process. The second mode applies when a new trial is perceived as having been produced by the same speaker as an immediately preceding trial. It is viewed as a form of extrinsic normalization, in which a listener's estimate of the formant-space is refined from the estimate used in the previous trial and applied to the identification of the current stimulus. Although a full account of the details of this process will require substantial additional research, the broad outlines seem rather clear.

Works Cited

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.). *Auditory analysis and perception of speech*. London: Academic Press. 103-113.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America*, *71*(4), 975–989. doi:10.1121/1.387579

Creelman, C. D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, *29*, 655. doi:10.1121/1.1909003

Forster, K. I. & J. C. Forster. DMDX: A Windows Display Program with Millisecond AccuracyBehavior Research Methods Instruments and Computers, Vol. 35, No. 1. (2003), pp. 116-124.

Gerstman, Louis. 1968. Classification of self-normalized vowels. IEEE Transactions of Audio Electroacoustics AU-16:78-80.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. doi:10.1121/1.411872

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857. doi:10.1121/1.398894

Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. doi:10.1121/1.1908694

Macchi, M. J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *The Journal of the Acoustical Society of America*, *68*(6), 1636–1642. doi:10.1121/1.385219

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *15*, 676–684.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. doi:10.1037/0096-1523.33.2.391

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. doi:10.1121/1.397688

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. doi:10.1121/1.397861

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America*, *80*(5), 1297–1308. doi:10.1121/1.394433

Nearey T. M., and Assmann P. F. (2007). Probabilistic 'sliding template' models for indirect vowel normalization. In *Experimental Approaches to Phonology*, edited by M. J. Sole´, P. S. Beddor, and M. Ohala (Oxford University Press, Oxford), pp. 246–269.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, speech production, and linguistic structure* (pp. 113–134). Tokyo: OHM.

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. doi:10.1121/1.1906875

Rakerd, B., Verbrugge, R. R., & Shankweiler, D. P. (1984). Monitoring for vowels in isolation and in a consonantal context. *The Journal of the Acoustical Society of America*, *76*(1), 27–31. doi:10.1121/1.391114

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America*, *117*(1), 305. doi:10.1121/1.1828637

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100. doi:10.1121/1.393381

Verbrugge, R., Strange, W., & Shankweiler, D. (1974). What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America*, *55*(S1), S53–S54. doi:10.1121/1.1919793

Watkins, A.J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *The Journal of the Acoustical Society of America*, *99*, 3749.

Watkins, Anthony J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *The Journal of the Acoustical Society of America*, *96*(3), 1263–1282. doi:10.1121/1.410275

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, *93*(4), 2152–2159. doi:10.1121/1.406678

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural Bases of Talker Normalization. *Journal of Cognitive Neuroscience*, *16*(7), 1173–1184. doi:10.1162/0898929041920522

Chapter 4

Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices

4.1 Introduction

Since the first acoustic studies in the 1950's, variation in the acoustic properties of vowels of different speakers has typically been discussed in terms of their fundamental frequency (f0) and formant frequencies (FFs). The scaling of f0 and FF ranges has also figured prominently in parametric synthesis of voices simulating speakers of different sizes, genders and age groups (Klatt and Klatt 1990). Although the perception of f0 has been extensively studied, the perception of the acoustic characteristic associated with the range of formant frequencies produced by different speakers is not as well understood. In the sections that follow, a case will be made for the importance of this acoustic characteristic, that we will call formant-frequency scaling (or FF-scaling), in the listener's assessment of apparent speaker characteristics (i.e., the indexical characteristics of the speaker inferred by the listener), and the perception of vowel quality. Furthermore, we suggest that the importance of FF-scaling in both vowel perception and the determination of apparent speaker characteristics may explain the relationship between these processes observed in several previous experiments.

In the discussion below, we will be adopting the uniform scaling hypothesis as a working assumption. Uniform scaling proposes that a set of phonetically equivalent vowels produced by two speakers of the same dialect are (on average) relatable to each other by a single multiplicative parameter. Although there is some controversy about this in the literature (see Appendix 3), in practice it leads to reasonably good approximations of systematic speaker variability (Nearey 1978, Nearey and Assmann 2007, Turner et al. 2009). The

scaling parameter (i.e., FF-scaling) is related to speaker vocal-tract length and determines the relative scaling applied to the formant-pattern of a given vowel by the vocal tract of the speaker.

4.1.1 FF-scaling and apparent speaker characteristics

Because of their dependence on the anatomy of the speaker, the average f0 and FFs produced by a speaker co-vary with some prominent speaker characteristics. Men tend to have lower f0s than women, and children tend to have higher f0s than adults of the same gender so that f0 correlates strongly with speaker height across all speakers (Hollien et al. 1994). The average FFs produced by a speaker will be most strongly determined by that speaker's vocal tract length, with longer vocal tracts producing lower FFs overall, and shorter vocal tracts producing higher FFs overall (Fant 1960). There is a strong positive correlation between speaker height and speaker vocal-tract length (Fitch and Giedd 1999) so that, in general, larger speakers have lower FF-scalings overall than smaller speakers (Lee et al. 1999; Peterson and Barney 1952). Consequently, the f0 and FFs of a vowel represent two potentially different streams of information arising from two acoustically distinct origins, each of which may be used by listeners to estimate speaker characteristics, such as height or gender.

Speakers may be divided into four general speaker classes based on two dichotomies: child vs. adult and male vs. female. If speakers are sorted to fit into one of these categories, then the average f0 and FF-scaling differences between speaker classes can be quite large. For example, an automatic classifier can predict the gender of an adult speaker with up to 98% accuracy using only information regarding the FF-scaling and f0 that characterize that voice (Hillenbrand and Clark 2009). However, the correlation between speaker height and voice characteristics (FF-scaling and f0) within a single class (e.g., adult males) is unreliable, particularly for adult speakers who have reached a stable height. There is no significant correlation between adult speaker height and average f0 after controlling for gender (Hollien et al. 1994, Gonzalez 2004, Lass and Brown 1978, Collins 2000, van Dommelen and Moxness 1995). It has similarly been reported that there is no significant correlation between adult

speaker height and FF-scaling after controlling for gender (Collins 2000, van Dommelen and Moxness 1995), or that the correlation is weak[15] (Gonzalez 2004).

Given that the relationship between the acoustic properties of the vowels produced by a speaker and that speaker's height is weak within a speaker class, it is not surprising that listeners are not able to accurately estimate speaker height based on a speaker's f0 and FF-scaling, when speaker class is controlled for, for example, by presenting listeners with speech from adult speakers only (van Dommelen and Moxness 1995, Collins 2000, Rendell et al. 2007). Despite the inability of listeners to arrive at *veridical* estimates of speaker size based on speech samples, listeners typically arrive at *consistent* judgments regarding a speaker's size, both within and across listeners (von Dommelen and Moxness 1995, Collins 2000, Smith and Patterson 2005, Rendell et al. 2007).

The manner in which listeners estimate speaker height has been investigated by presenting listeners with speech sounds that vary in terms of f0 and FF-scaling, but with a fixed phonetic content, and asking listeners to assess the absolute or relative heights of speakers. This has been done using synthetic vowels (Fitch, 1994) and modified natural-speech (Ives et al. 2005, Smith and Patterson 2005, Smith et al. 2005, Rendell et al. 2007). Results indicate that these judgments are informed by jointly considering the FF-scaling and f0 of a voice (Fitch 1994, van Dommelen and Moxness 1995, Smith and Patterson 2005), where progressively lower FF-scalings and/or progressively lower f0s suggest a progressively larger speaker.

Most listeners are familiar with the concept of pitch and it is known that they can make overt judgments of pitch that relate to the relative f0 level of

[15] Lack of significance could be in part due to the reduced power of tests based on small number of observations compared to the full sample. This is at least partly due to the restricted ranges used when considering correlations between acoustic characteristics of speech and the physical qualities of the speaker only for a restricted class of speakers. By restricting the range of a predictor when the error in the response variable remains constant, the correlation between two variables will become weaker (Bland and Altman 2011, Sackett and Yang 2000.). In the most extreme example, the correlation between the acoustic properties of voices and the heights of men who are all the same height will necessarily be zero.

different voices (Honorof and Whalen, 2005). It is not clear, however, whether there exists any separable perceptual dimension that corresponds closely to FF-scaling that listeners might learn to report. Since this putative perceptual dimension[16] has no name that we know of, we will refer to it tentatively as the perceptual FF-scale estimate, or pFF-scaling, to keep it distinct from the acoustic FF-scaling used to create the stimuli used in the experiment to be outlined below.

To date, experiments involving listener responses to variations in the FF-scaling of voices have focused on the estimation of speaker characteristics (e.g., gender, body size), which are determined by jointly considering voice f0 and FF-scaling. For example, a common methodology (Fitch 1994, Smith and Patterson 2005) involves creating a set of stimuli with fixed phonetic content, which span an f0 x FF-scaling space (as in Figure 4.2). Listeners are then presented with these stimuli in a random order and, for each trial, are asked to estimate some speaker characteristic, for example, the speaker's height or gender. By comparing the rated heights of voices at different points within an f0 by FF-scaling space, researchers may investigate the relative contribution of each cue to such judgments via linear regression. Although this methodology can shed light on the manner in which speaker characteristics are determined by jointly considering voice f0 and FF-scaling, they cannot provide information about listeners' use of any perceptual dimension or mechanism that specifically follows physical variation in FF-scaling as such.

[16] As far as we have been able to determine, this perceptual property has no specific name in either psychophysical or musical terminology, although it appears to bear some relation to some subdivisions of the German Fach system of classification of operatic voices. Such a perceptual property might correspond to the scale-dimension of what Patterson and colleagues propose is a Mellin(-like) transform performed by the peripheral auditory system that segregates information related to vocal-tract length from information relation to vocal-tract configuration. In Section 4.1.2, we suggest that pFF-scaling might be a kind of derived perceptual property, which is determined when a listener establishes a speaker-dependent frame of reference. The location of that frame of reference is indexed by a single scalar value, analogous to Ψ in Nearey and Assmann's (2007) sliding template model, and the parameter a seen in Equation 1 presented in Turner et al (2009, p. 2377) .

For example, consider two voices with the same f0 and source characteristics, one of which has a lower FF-scaling than the other. If one listener reports hearing a male for the low FF-scaling voice, and a female for the high FF-scaling voice, it is reasonable to infer that they are responding to a change in voice FF-scaling. However, if a second listener reports that both voices appear to represent male speakers, this does not entail that the listener fails to notice the difference in FF-scaling. Rather, the second listener may have a higher threshold for a change in apparent speaker gender, or they may attribute the change in FF-scaling to a change in size-within-gender or any number of factors (including, for some formant patterns at least, differences in vowel quality whether categorical or graded). In short, the collection of judgments of apparent speaker characteristics does not allow researchers to directly investigate the perception of FF-scaling or its putative perceptual counterpart pFF-scaling. As discussed below, if listeners are able to provide perceptual judgments that correlate well with FF-scaling, such judgments could be a valuable source of information in the evaluation of perceptual theories related to vowel-normalization.

4.1.2 FF-scaling, normalization and vowel perception

Several theories of human vowel perception involve the estimation of a speaker-dependent formant-space as a frame of reference used to interpret the vowels produced by a speaker (Joos 1948, Ladefoged and Broadbent 1957, Ainsworth 1975, Nearey 1978, Nearey 1989, Nearey and Assmann 2007). The speaker-dependent formant-space need only be detailed enough so that a listener knows roughly what FFs to expect for a given vowel category when produced by that speaker. The listener then identifies vowels by considering the FFs of a vowel sound relative to expected FFs for each vowel category, rather than by considering the FFs in an absolute manner. This general hypothesis is typically referred to as *speaker normalization*. To the extent that variation in formant-spaces across speakers can be accounted for by a single parameter (i.e., FF-scaling), the process of speaker normalization can be thought of as centering

around the estimation of an appropriate FF-scaling with which to identify vowels produced by that speaker.

This insight underlies the log-mean normalization method proposed in Nearey (1978). It has been used routinely for decades in sociophonetic studies by Labov and his colleagues, where it has been found to be effective for preserving relatively subtle systematic differences between dialects and sociolects while largely removing effects of vocal tract length (Labov, Ash, and Boberg, 2006, p. v). This method calculates the log-mean FF produced by a speaker across their entire vowel system, a measure which should be strongly correlated with speaker FF-scaling, and subtracts this value from the log-transformed formant frequencies produced by a speaker. In effect, this method centers the vowel spaces of different speakers along the primary axis of variation between speakers (i.e., ln F1 = ln F2, see Appendix 3) and, consequently, allows variation in FFs to be interpreted more directly as evidence of differences in vowel quality (as opposed to simply being a result of differences in speaker vocal-tract length).

Consider Figure 4.1, which presents the Peterson & Barney (1952) vowel data. In this figure FFs have been normalized using the log-mean method of Nearey (1978). As seen in Figure 4.1, this process greatly reduces the between-category overlap between vowel categories relative to the raw FFs (presented in Figure A3.1 in Appendix 3). Furthermore, the major axes of the ellipses representing the different vowel categories are no longer primarily aligned with the ln F1 = ln F2 axis as they are for unnormalized data (see Appendix 3). In fact, whereas variation along this axis accounted for 80.6% of the variance in FFs in the unnormalized FFs (a ratio of nearly 4/1), after normalization variation along this axis accounts for only 52.9% of variation, on average indicating an essentially equal distribution in variation along ln F1 = ln F2 and the orthogonal axis.

Although a speaker-dependent FF-scaling estimate may play an important role in vowel perception, the listener does not have direct access to the speaker's true FF-scaling, and must estimate this value. Both Nearey and Assmann (2007) and Turner et al. (2009) have emphasized that since the uniform scaling hypothesis entails that productions between speakers of the same vowel differ by

a single multiplicative parameter (i.e., FF-scaling), identifying a vowel sound will yield an estimate of the speaker-specific parameter (i.e., pFFscaling), since listeners may infer the speaker's FF-scaling given the observed formant frequencies. This is analogous to the manner in which identifying a visual object of a known physical size yields an estimate of its distance from the observer. In this view of vowel perception, the speaker-dependent FF-scaling estimate, pFF-scaling, might be thought of as a derived perceptual property, which a listener constructs in establishing a speaker-dependent formant-space with which to interpret a speaker's vowels.

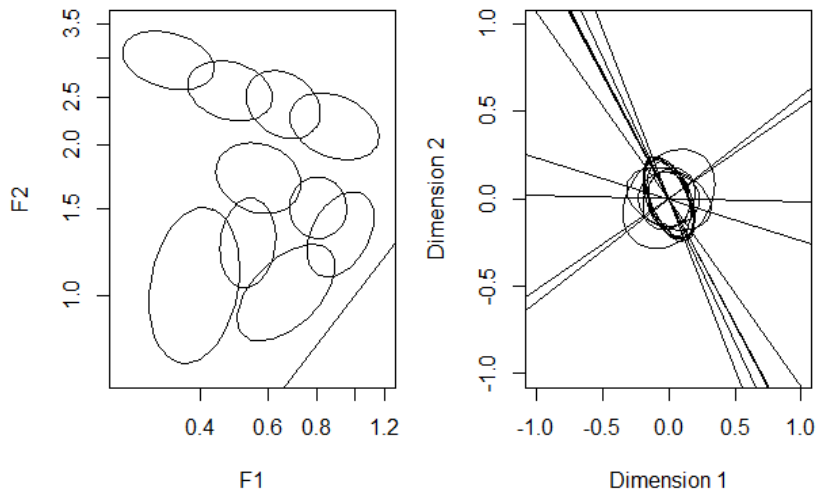

Figure 4.1. *In the left panel, ellipses enclosing two standard deviations of the Peterson and Barney (1952) vowels are presented. Vowels have been normalized-using the log-mean normalization method of Nearey (1978). F1 and F2 are presented as the ratio of each formant frequency to the geometric mean F1-F2-F3 frequency produced by each speaker across their whole vowel system. The line is a line parallel to ln F1 = ln F2. In the right panel, all formant frequencies have been log-transformed and centered within-category so that vowel-category means are at the origin. All points have been rotated 45 degrees clockwise so that the ln F1 = ln F2 line is now parallel to the x-axis (Dimension 1), while Dimension 2 represents the orthogonal axis. The lines indicate the major axes of the vowel-category ellipses, which no longer vary primarily along the ln F1 = ln F2 axis (Dimension 1).*

4.1.3 FF-scaling, vowel perception and apparent speaker characteristics

Because of the potential importance of FF-scaling estimates in human vowel normalization, the ability to collect them from listeners may help clarify unresolved issues in the study of speech perception. For example, previous studies have found that vowel quality shifts can be induced by manipulating vowel f0, or the f0 of a preceding carrier phrase (Miller 1953, Fujisaki and Kawashima 1968, Slawson 1968, Nearey 1989, Johnson 1990). Similar effects have been observed by pairing vowel sounds with male or female faces (Glidden and Assmann, 2004) or simply by telling listeners that the speaker is of a certain gender (Johnson et al., 1999). Johnson (1990, 1999, 2005) has suggested that f0 affects vowel quality primarily indirectly, by affecting apparent speaker characteristics, rather than by being directly involved in the specification of vowel quality.

In terms of a general theory of speaker normalization, f0 is expected to affect perceived vowel quality primarily by informing the speaker-dependent formant-space used by the listener to interpret the vowels of a speaker. Apparent speaker gender is expected to affect perceived vowel quality in a similar manner. For example, if a vowel is presented with a high pitch, a listener may assume that the speaker is a female and may assume a formant-space appropriate for a female speaker. If a vowel with the same FFs were presented with a low pitch, the listener may assume a male speaker, and a formant-space appropriate for a male, which may lead to differences in perceived vowel quality. This may be contrasted with the direct effect of a change in F1, for example, which would be expected to result in a change in vowel quality even within-speaker.

Barreda and Nearey (2012a) report the results of an experiment that offers strong support for Johnson's hypothesis. Listeners were presented with a series of vowels that differed in their FFs and f0 and, for each trial, were asked to report vowel quality and two apparent speaker characteristics. The speaker characteristics they were asked to report were speaker gender (male or female) and speaker size (using a continuous scale that they were instructed to use as they saw fit). Results indicate that although f0 can exert a strong influence on perceived vowel quality, this effect is greatly diminished (but still significant) if

apparent speaker characteristics are accounted for. This was taken as an indication that although f0 is strongly related to perceived vowel quality, its effect is mostly achieved by suggesting apparent speaker characteristics to the listener. Furthermore, apparent speaker gender had a significant effect on perceived vowel quality, and apparent speaker size (controlling for gender) had a marginally significant effect[17] on vowel quality, even after controlling for the acoustic characteristics of the vowel sound.

Although experiments such as Johnson (1990), Johnson et al. (1999), Glidden and Assmann (2004) and Barreda and Nearey (2012a) used speaker characteristics such as speaker gender to investigate the process of speaker normalization, none of these authors suggest that speaker gender is directly involved in the specification of vowel quality in the same way that the formants are. Rather, these experiments might be interpreted as using apparent speaker characteristics as surface variables to investigate the latent variable of interest, the FF-scaling estimate for a voice on the part of the listener. Because of the strong and consistent association listeners make between FF-scaling and perceived speaker size and gender (outlined in Section IA), experimenters might reasonably infer that if listeners indicate that a speaker is an adult male, they will also expect a relatively lower FF-scaling than if the speaker were an adult female. Thus, controlling for apparent speaker characteristics, as in Barreda and Nearey (2012a), can be viewed as indirectly attempting to control for a latent estimated FF-scaling, while affecting apparent speaker gender as in Glidden and Assmann (2004) might be viewed as an attempt to influence implicit, listener-internal FF-scaling estimates.

A more direct approach to experiments investigating the direct and indirect effects of acoustic cues on vowel quality would be to collect overt FF-scaling

[17] A positive relationship was expected between perceived vowel quality and apparent speaker size, and 14 of 19 participants exhibited a positive relationship between the two variables. This corresponds to a one-tailed p-value of 0.0318 using a non-parametric sign test. However, a t-test of the same partial correlations finds that they are not significantly different from zero ($p = 0.3027$).

judgments from listeners in experiments designed to investigate specific questions. If this could be done, researchers would not need to rely solely on speaker characteristics that, although they may strongly co-vary with speaker FF-scaling, may do so only in a complex, derivative way. Furthermore, specific hypotheses about the possible role of FF-scaling estimates in vowel perception could be tested in a more direct manner.

4.1.3 Rationale for the Current Experiment

In the previous sections we have established that the formant patterns produced by speakers of different sizes vary primarily in terms of a single, multiplicative parameter, which we refer to as FF-scaling. Because of its strong relationship to speaker vocal-tract length, this acoustic characteristic is closely related to salient apparent speaker characteristics such as size and gender. Listeners may take advantage of this co-variation, and use FF-scaling information to infer apparent speaker characteristics from the speech signal. We have outlined a case for the potential centrality of information related to speaker FF-scaling in human vowel perception in terms of a general process of speaker normalization. Finally, we have suggested that the effect of some apparent speaker characteristics on perceived vowel quality may occur by means of influencing the listener's speaker-dependent FF-scaling estimate.

Although the line of reasoning summarized in the previous paragraph has extensive experimental and theoretical support, the perception of speaker FF-scaling is not well understood. Given that our position on the process of vowel perception centers around a speaker-dependent FF-scaling estimate, it is incumbent on us to demonstrate that listeners are able to identify voices that differ according to this acoustic characteristic, and to investigate the nature of a possible pFF-scaling perceptual dimension.

Despite the potential usefulness of obtaining voice FF-scaling estimates from listeners, no previous experiment has focused on training listeners to directly report this property. The purpose of this experiment is to investigate the extent to which listeners can learn to distinguish and identify voices that vary in both average f0 and FF-scaling. The experiment to be outlined here adopts a similar

stimulus design to that employed in Fitch (1994) and Smith and Patterson (2005), where listeners are presented with a series of stimuli that span an f0 x FF-scaling space but have a fixed phonetic content. However, instead of a rating-scale judgment of a specific speaker characteristic, listeners are trained to provide absolute identifications of each voice presented from a discrete set of alternatives in a two-dimensional display corresponding to an f0 x FF-scaling space. In doing so, listeners will provide what can be viewed as estimates of voice f0 and voice FF-scaling independently for each dimension[18], rather than providing a measure (such as judged size or gender) that is likely to involve joint consideration of the two properties.

This experiment also seeks to investigate the feasibility of collecting FF-scaling judgments from listeners in varying experimental conditions. Future experiments investigating the manner in which listeners estimate voice FF-scaling may require listeners to report voice f0, or they may require listeners to disregard it, depending on the specific question being addressed. To investigate whether disregarding stimulus f0 results in a significant change in the consistency with which listeners report voice FF-scaling, the ability of listeners to report voice FF-scaling will be tested in two conditions. In the first of these, listeners will be asked to report FF-scaling and f0 for each trial. In the second condition, listeners will be asked to report FF-scaling only, and disregard stimulus f0.

There are three general possible outcomes, each of which has different implications for the manner in which human listeners respond to and isolate the FF-scaling of a voice, and for the nature of an acoustic quality such as pFF-scaling. The first possible outcome is that listeners are not able to do this and perform no better than chance in either of the testing conditions. This outcome would be problematic given that listeners have been found to respond to FF-

[18] In our analysis we will assume listener's judgments are really separated into these two components at the time of choice. However, even if listeners were instead memorizing a discrete set of individual voices, the systematic correspondence of their choices to the FF-scaling and f0 dimensions would at least provide evidence that the 'perceptual speaker space' is organized in in a way that includes a subspace that is effectively a near projection of these two dimensions.

scaling changes in determining apparent speaker characteristics. This outcome might suggest that listeners' representations of voice characteristics are not organized along dimensions related to FF-scaling; that the training paradigm was fundamentally flawed in some way; or, finally, that the task was too difficult given the relatively short training sequence.

The second possible outcome is that listeners are able to report their judgments of FF-scaling with a consistency and accuracy (that is, the judgments are strongly correlated with the physical FF-scaling of the stimuli), and that these judgments are made independently of stimulus f0. This outcome would be predicted based on the work by Irino and Patterson (2002), Smith et al. (2005), and Turner et al. (2006), which have all suggested that the peripheral auditory system processes sounds at an early level, and that this processing segregates information regarding the size of the vocal tract from information regarding the particular configuration of the vocal tract during articulation. The output of this process is expected to be directly available to the listener (which would suggest relatively high performance), and FF-scaling identification should not be influenced by f0.

The third possible outcome is that listeners are able to report FF-scaling with good level of accuracy and consistency, but that these judgments are influenced by stimulus f0. This outcome would be predicted by processes similar to Method 6 of the Sliding Template Model (Nearey and Assmann 2007), which estimates speaker FF-scaling on the basis of the joint distribution of f0 and FF-scaling between speakers, and the relative fit of the observed FFs to those expected for each vowel category. Importantly, only a main effect of f0 on reported FF-scaling is predicted, where a higher f0 should result in a higher reported FF-scaling. This predicted outcome will be shared by any proposed normalization method which seeks to exploit the covariance between FF-scaling and f0 between speakers to estimate speaker FF-scaling based on f0 (although specific models may predict more complicated patterns of relationships between f0 and reported FF-scaling).

4.2 Methodology

4.2.1 Participants

Listeners were 71 students from the University of Alberta drawn from a participant pool in which undergraduate students take part in experiments in exchange for partial course credit. All participants were students taking an introductory level, undergraduate linguistics course. Before beginning the experiment, all participants filled out a questionnaire in which they indicated their age, gender, native language, any other languages they spoke, and the amount of formal musical training they had received (measured in years). This background information was collected because we thought that prior musical or language experience might influence listeners' ability to perform the experimental tasks successfully. Our reasoning is discussed further in Section 4.3.

4.2.2 Stimuli

The stimuli consisted of vowel pairs with formant patterns appropriate for the sequence [i æ] (in that order, separated by a pause) spoken by a single speaker. These were constructed to simulate the voices of 15 different synthetic speakers. The vowels associated with these voices varied on the basis of three factors: f0 step, FF-scaling step, and the difference in FF-scaling between adjacent FF-scaling steps (this difference will be referred to as ΔFF–scale). FF-scaling level and f0 level were within-subjects factors, so that each listener was presented with voices at each combination of f0 and FF-scaling steps (3 f0 steps x 5 FF-scaling steps). However, ΔFF–scale was a between-subjects factor, so that each listener was only ever presented with voices at a single ΔFF–scale level.

The FFs of vowels representing an FF-scaling step were determined by increasing all of the FFs of the previous step by a fixed percentage (i.e. by a single multiplicative scale factor). The size of the percentage increase between adjacent FF-scaling steps was determined by the ΔFF–scale level. Four different FF-scaling increments were used (7%, 8%, 9%, 10%), resulting in four groups of listeners. For example, for the stimuli for the 9% ΔFF–scale level, the FFs of the vowels of the second FF-scaling step were determined by increasing all of the FFs

of the first FF-scaling step by 9%. The FFs of the vowels for the third FF-scaling step were then increased by a further 9% relative to those of the second step (18.81% relative to the first FF-scaling step), and so on.

Table 4.1. *Initial f0 levels for all conditions. Formant frequencies provided are those used for the lowest FF-scaling step vowels in all conditions, corresponding to formant frequencies appropriate for a typical adult male.*

	Low	**Med.**	**High**
f0	110	177	270

	F1	**F2**	**F3**
i	280	2148	2755
æ	717	1497	2318

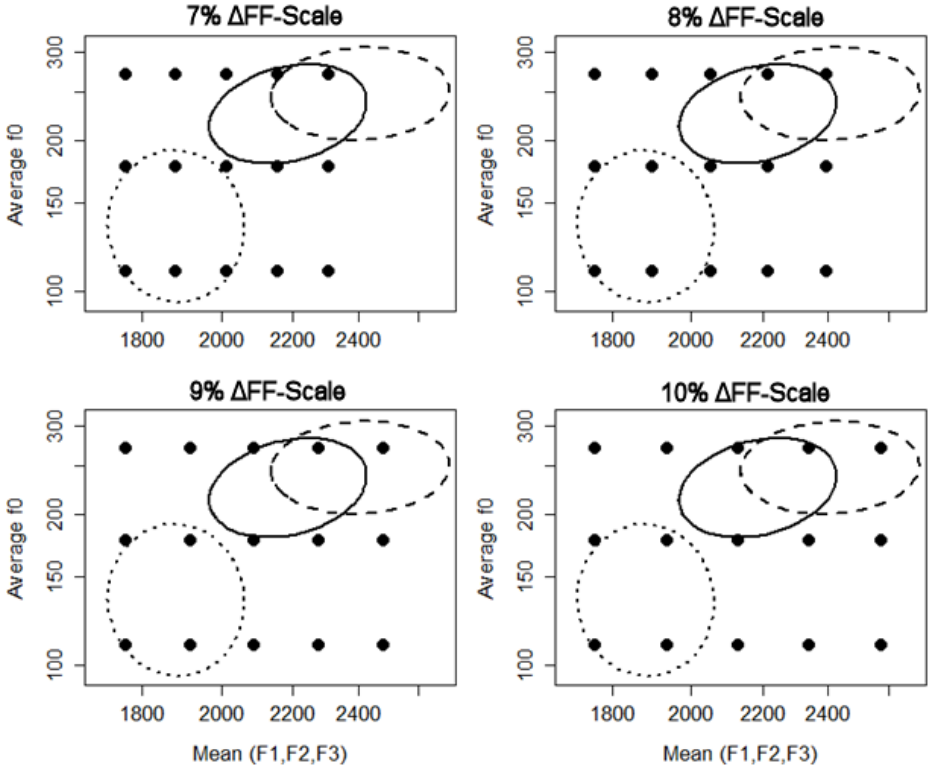

Figure 4.2. *The x axis indicates the mean of the first three formant frequencies for productions of /i/. Ellipses enclose two standard deviations of the distribution of real voices from data collected by Hillenbrand et al. (1995). Ellipses indicate the distribution of voices of adult males (dotted line), adult females (solid line) and children (broken line). The locations of stimulus voices at each ΔFF–scale level are indicated by the filled points.*

It is worth noting that the FF-scaling differences used in the construction of the stimuli for this experiment (7%, 8%, 9%, 10%) are close to the estimated just noticeable difference for FF-scaling, estimated to be 7-8% by Smith et al. (2004) and 4-6% by Ives et al. (2005). In both cases, just noticeable differences were estimated using a two-alternative, forced-choice methodology.

Each vowel of the [i æ] stimulus pair was 200 ms in length, and these were separated by 125 ms of silence. Table 4.1 presents the initial values for each of the three f0 steps. For every stimulus, f0 decreased linearly by 10% from the beginning to the end of the vowel. F0 levels were the same for all ΔFF–scale levels. Table 4.1 also provides the FFs used for the first (lowest-frequency) FF-scaling step for all ΔFF–scale levels. These values were set based on average productions of the same vowels produced by adult male native speakers of the regional dialect. For both vowels, F4 was set at 3375 Hz and each formant above F4 was 1000 Hz higher than the last, up to the tenth formant. Vowels were synthesized with a variable sampling rate so that the Nyquist frequency fell halfway between the tenth formant and the expected frequency of the eleventh formant given the spacing between formants. The inclusion of higher formants, and the variable sampling rate, were undertaken to avoid inappropriate spectral levels that can readily result when there is an uneven distribution of formants near the Nyquist frequency (See Nearey 1989, Appendix B, for a discussion of some of the issues involved). All vowels were then re-sampled at 22050 Hz. Figure 4.2 compares the location of the synthetic voices used in this experiment, for each ΔFF–scale level, to a range of real voices plotted on an f0 x FF-scaling space.

4.2.3 Procedure

A training game reminiscent of the 'concentration' or 'memory' card game was created to train participants to report FF-scaling independently of f0. This game was played on a computer using a specially-designed graphical user interface. The game board contained 15 boxes arranged in three rows of five. Each of these boxes was associated with a single voice throughout each participant’s experimental session. Voices in the same row had the same f0 while voices in the same column had the same FF-scaling. Voice f0 increased from top

to bottom across rows while voice FF-scaling increased from left to right across columns (in fact, the stimulus voices were arranged on the board in the same manner that they are arranged in Figure 4.2). Before beginning the game, participants completed an introductory task in which they were familiarized with all voices. Participants were told that the pitch of voices would increase from bottom to top and that voices differed from left to right in terms of 'voice size', which they were told was closely related to speaker size.

The general procedure during the training game was that participants were presented with vowels produced by one of the voices on the board and were asked to indicate the position of the voice within the board by clicking on the box that was associated with it. By locating the voice on the board, participants were, in effect, reporting the FF-scaling and f0 levels for the stimulus voice. The game consisted of a series of 11 levels of increasing difficulty. Difficulty was increased between levels by increasing the number of candidate voices available to listeners during each trial. For example, initially listeners were asked to identify a voice from one of two candidates, while in later levels listeners were asked to identify a voice from among all voices in a row, or all voices in two rows. Buttons associated with voices that were candidates for selection in the session were colored blue. Buttons that were not to be considered for selection were the same grey color as the background of the board.

The procedure in each level was as follows: For a trial, listeners were played the vowels [i æ], produced by a single voice. These vowels were always presented in the same order and were separated by 125 ms of silence. Listeners were allowed to replay the vowels as many times as they liked by clicking on a button marked 'replay'. Listeners then had to indicate the location of the voice on the board by clicking on one of the blue buttons. When listeners answered correctly, the next pair of vowels played after a 1 second pause, and the process continued until all candidate voices were identified three times each. Voices were presented in a randomized condition, blocked by repetition.

When participants answered incorrectly, the game entered into a special game mode designed to provide the user with feedback, and an opportunity to

improve their performance by listening to the voices on the board. In this mode, the correct location of the voice the listener had just heard was indicated by a green box. The box that had been incorrectly selected by the listener was indicated by a red box. Listeners were allowed to listen to all available voices as many times as they liked by clicking on the boxes associated with different voices. When a listener was finished using error mode, they clicked on a button marked 'resume', after which the next voice in the round was presented after a one second pause.

Longer-term feedback was provided to listeners via a message across the top of the game board, which informed listeners of the percent of trials they had identified correctly within a given level and of the percentage of trials in which they had been within one step at most, in both f0 and FF-scaling level, of the correct box. When a level was completed, listeners moved on to the next level in the game by clicking on the button marked ‘resume’. The next level would not begin until the listener clicked on this button. All listeners took part in experimental sessions of a maximum of one hour in length.

After completing all levels of the training game, listeners performed two experimental tasks. In the first task listeners were asked to identify a voice from among all 15 candidate voices by indicating its f0 and FF-scaling level. This task will be referred to as the Two-factors task. Listeners identified each voice three times, for a total of 45 trials per participant for this task. The Two-factors task should give the best indication of the ability of listeners to separate FF-scaling and f0 information and to report each independently. For the second task, listeners were again asked to identify stimuli from among all candidate voices, however, for this task listeners only had to indicate stimulus FF-scaling level and ignore f0 (this will be referred to as the FF-only Task). For this task, only the middle row of response buttons were visible to the listener so that listeners only had the option of reporting FF-scaling. Again, listeners identified each voice three times, for a total of 45 trials per participant for this task. This task was intended to compare the ability of listeners to identify stimulus FF-scaling when listeners are asked to

report f0 and when they are asked to ignore f0. All listeners performed the Two-factors Task before the FF-scaling only Task.

4.3 Results

The performance of different listeners was expected to vary as a result of two main classes of characteristics. The first of these is the different scaling-factor increments (ΔFF-scale) used to create the synthetic voices. Since larger ΔFF-scales increase the acoustic difference between adjacent FF-scaling levels (i.e., horizontally adjacent voices on the board), it was expected that ΔFF-scale level would affect identification rates, with lower values resulting in worse performance. This is a between-subjects factor in the statistical design and can be dealt with directly as such.

The second class of characteristics expected to affect listener performance is the differences in ability that participants may have had before beginning the training, or the different rates at which participants might learn to independently report the two aspects of voice quality being investigated here. Although no direct measure of these differences is available independently of the experimental results, it was expected that three additional characteristics that relate to listeners' background experience could serve as covariates that reflected these differences in ability.

ΔFF-scale	7	8	9	10
Total Listeners	18	18	18	17
English Native Speakers	17	14	15	15
Fluent in a Tone Language	5	4	4	4
Musically Trained	7	9	8	6

***Table* 4.2.** *Distribution of some listener characteristics among different ΔFF-scale groups.*

The first of these covariates is native language, where the performance of native speakers of English might differ from that of non-native speakers. For example, non-native speakers might have more difficulty processing the categorical vowel information and might be operating under a greater cognitive load than native speakers. The second covariate is fluency in a tone language. Seventeen participants were fluent in a tone language. These speakers may have had an advantage in identifying pitch levels or in separating pitch and FF-scaling information relative to speakers without knowledge of a tone language. The final covariate was the number of years of formal musical training a listener had received (including zero for listeners who had received no musical training). In pilot tests of the training program, a listener who was a trained musician performed considerably better than any other listener. It was anticipated that formal musical training might also help listeners learn to separate the f0 and FF-scaling information of sounds independently and thus might affect performance. The distribution of these characteristics among listeners in different ΔFF-scale groups is presented in Table 4.2.

4.3.1 Identification of voice f0 and FF-scaling

4.3.1.1 Performance for the Two-factors Task

Identification rates were found for correct labeling of f0 level, correct labeling of FF-scaling level and correct absolute identification (where both factors were correctly labeled), individually for each participant (n = 71). Performance was high overall with an average of 79.4% f0 identifications (min = 31%, max = 100%, sd = 15.3%), 40.1% correct FF-scaling identifications (min = 15.5%, max = 71%, sd = 12%), and 33.6% correct absolute identifications of both characteristics simultaneously (min = 6.7%, max = 71%, sd = 13.7%). All three mean values were considerably higher than what would be expected given chance performance (33%, 20% and 6.7% respectively). There was a moderate positive correlation between correct identification rates for f0 and FF-scaling within-listeners; listeners who identified f0 at a higher rate also identified FF-scaling at a higher rate [$r = 0.44$, $t(69) = 4.1$, $p = 0.0001$].

Listener performance was expected to be affected by the between-subjects factor ΔFF-scale. In addition, the covariates reflecting listeners' background experience were also expected to influence performance level. In order to test which of these characteristics had a significant effect on performance on the Two-factors Task, a regression analysis was carried out on the within-participant, correct absolute-identification rates. The predictor variables were the between-subjects factor ΔFF-scale (7%, 8%, 9%, 10%), the binary indicator variables native language (English vs. non-English), tone language fluency (fluent vs. not fluent), and the level of musical instruction, coded as a continuous covariate (in number of years of instruction, including zero for listeners who had received no instruction).

None of the effects reached significance, except the effect of musical training [$F(1,64) = 16.8$, $p = 0.0001$]. Surprisingly, the main effect for ΔFF-scale did not even approach significance [$F(3, 64) = 1.4$, $p = 0.25$]. Thus, listeners in the 7% ΔFF-scale group scored about as well as those in the 10% ΔFF-scale group, 37% and 35% correct absolute identifications respectively. A parallel analysis of variance was carried out on the marginal correct identification rates for voice f0 and FF-scaling. These analyses revealed a similar pattern of results with the only significant main effect being for musical training for correct identification of f0 [$F(1,64) = 17.8$, $p < 0.0001$] and FF-scaling [$F(1,64) = 9.9$, $p = 0.0025$].

4.3.1.2 Performance for the FF-scaling only Task

Since only information regarding FF-scaling estimates was collected for the FF-scaling only Task, all references made to correct identification rates refer to FF-scaling identification alone. Once again, correct identification rates were found individually for each participant ($n = 71$). Performance was high overall, with an average correct FF-scaling identification rate of 40.6% (min = 13.3%, max = 64%, sd = 11.8%), which is very close to the 40.1% correct FF-scaling identification rate for the Two-Factor Task.

A regression analysis was carried out in which FF-scaling identification rate was the dependent variable. Once again, the predictor variables were the between-subjects factor ΔFF-scale (7%, 8%, 9%, 10%), the binary indicator

variables native language (English vs. non-English), tone language fluency (fluent vs. not fluent), and the level of musical instruction, coded as a continuous covariate (in number of years of instruction). The same pattern of effects was found as in Two-Factor Task, with only musical training [$F(1, 64) = 9.5$, $p = 0.0030$] being a significant predictor of participant performance.

Finally, in order to see if a listener's ability to identify voice FF-scaling was affected by whether they were also asked to report voice f0, a t-test was carried out on the individual, within-participant difference in FF-scaling identification across the two tasks. The mean within-listener difference in performance between the two task was 0.5%, a difference that did not reach significance [$t(70) = .44$, $p = .66$]. This indicates that voice FF-scaling estimation is similar in cases where listeners are asked to report voice f0 and in cases where they are asked to disregard it.

4.3.2 Information used in FF-scaling estimation

The FF-scaling indicated by the listener in response to a trial will be referred to as judged FF-scaling, as opposed to the veridical stimulus FF-scaling level present in each stimulus. Judged FF-scaling is expected to correlate strongly with the listener-internal pFF-scaling perceptual dimension. Consequently, the most important determiner of judged FF-scaling was expected to be stimulus FF-scaling. If listeners were performing this task using only information from the FFs of a vowel pair to determine the FF-scaling of the voice that produced them, stimulus FF-scaling would be the only significant predictor of judged FF-scaling, with no role for stimulus f0. On the other hand, a significant main effect for f0 may indicate a process of FF-scaling estimation such as Method 6 of the Sliding Template Model (Nearey and Assmann, 2007) where f0 may bias FF-scaling estimates. We know of no theory that would predict a significant interaction between f0 and FF-scaling in the determination of FF-scaling estimates.

The relationship between judged FF-scaling and stimulus f0 and FF-scaling was investigated using ordinal logistic regression. Models of this kind allow one to investigate the classification of stimuli into a sequence of discrete, ordinal categories based on a given number of explanatory variables. In this case,

the dependent variable was the judged FF-scaling provided by the listener for each trial. Judged FF-scaling steps were coded as one through five, where higher numbers indicated higher FF-scaling ratings (and higher average FFs for a voice). Stimulus FF-scaling was coded as a centered covariate, while stimulus f0 steps were coded using dummy variables, where the lowest f0 step acted as the reference group. This coding allows for a linear relationship between stimulus and judged FF-scalings, as well as for stimulus f0 levels to results in shifts in judged FF-scaling. The interaction between these two terms allows for the possibility that stimulus FF-scaling had a different linear relationship with judged FF-scaling at different levels of stimulus f0.

A model was fit to the data collected for each participant independently, and this was carried out separately for the data from each of the two tasks performed (Two-factors Task and FF-scaling only Task). Significance testing was then carried out on the coefficients found for each listener, for each task, to investigate the effects of each predictor on judged FF-scaling (Gumpertz and Pantula 1989).

For the Two-factors Task, stimulus FF-scaling was a highly significant predictor of judged FF-scaling [$F(1,70) = 77.9$, $p < 0.0001$]. As expected, there was a positive relationship between stimulus FF-scaling and judged FF-scaling. The main effect for f0 did not approach significance [$F(2,69) = 0.38$, $p = 0.68$]. However, the interaction between stimulus f0 and stimulus FF-scaling was significant [$F(2,69) = 8.79$, $p = 0.0004$].

The interaction between stimulus f0 and FF-scaling may be decomposed by stimulus f0 level. Since the lowest f0 step was used as the reference group, these interactions indicate whether the linear relationship between stimulus and judged FF-scaling differed significantly at the second or third f0 steps relative to the relationship observed for the lowest f0 step. When considered in this way, only the interaction between the second, intermediate f0 level and stimulus FF-scaling reaches significance [$t(70) = -3.08$, $p = 0.0029$]. The interaction is negative, resulting in a decrease in the slope relating stimulus FF-scaling to judged FF-scaling. Since the dependent variable representing stimulus FF-scaling

was centered, the decrease in slope indicates that responses tended to gravitate towards the middle of the FF-scaling response space for the middle f0 level more so than for the high and low f0 levels.

For the FF-scaling only Task, there was a very strong positive relationship between stimulus FF-scaling and judged FF-scaling [$F(1,70) = 55.8$, $p < 0.0001$]. Unlike for the Two-factors Task, stimulus f0 [$F(2,69) = 16$, $p < 0.0001$] had a significant (main) effect on judged FF-scaling. The effect of each of the stimulus f0 levels on judged FF-scaling was positive, indicating that higher stimulus f0s were associated with higher judged FF-scalings. The interaction between stimulus FF-scaling and stimulus f0 was also significant [$F(2,69) = 12.7$, $p < 0.0001$]. When decomposed by stimulus f0 level, this interaction showed a similar pattern as that observed for the Two-factors Task in that only the interaction between the second f0 level and stimulus FF-scaling reached significance [$t(70) = 2.66$, $p = 0.0096$]. Once again, this interaction was negative indicating a decrease in the slope relating stimulus FF-scaling to judged FF-scaling.

Two-factors Task			
Term	**df**	**Sum of Squares**	**% Var. Exp.**
FF-S	1	2244.2	35.6
f0	2	32.3	0.5
FF-S x f0	2	7.6	0.1
Residual	--	4023.2	63.8

FF-scaling only Task			
Term	**df**	**Sum of Squares**	**% Var. Exp.**
FF-S	1	1980.4	31.6
f0	2	242.2	3.9
FF-S x f0	2	8.9	0.1
Residual	--	4034.3	64.4

Table 4.3. *Sum of squares and percent of variance explained of judged FF-scaling explained by stimulus FF-scaling (FF-S), stimulus f0 (f0) and the interaction of the two.*

The significant effects for stimulus f0 in both models described above indicate that stimulus f0 does have an effect on judged FF-scaling. In order to get a rough estimate of the magnitude of these effects, two linear models were fit to the pooled data across all participants. This process was carried out independently for the results from the Two-factors Task, and those from the FF-scaling only

task. These models treated the response variable, judged FF-scaling, as a continuous variable. The independent variables were coded in the same manner as for the models outlined above. Table 4.3 presents the sum of squares and the percent variance explained by each of the explanatory variables for each of these models.

It is clear from the proportion of variance explained by stimulus FF-scaling that judged FF-scaling is most strongly determined by stimulus FF-scaling. In both the Two-factors Task and the FF-scaling only Task, stimulus f0 and the interaction between stimulus f0 and stimulus FF-scaling explain only a very small amount (0.1% to 3.9%) of the overall variance in judged FF-scaling. These results indicate that the significant effect of stimulus f0 on judged FF-scaling, as well as the significant interaction between stimulus f0 and stimulus FF-scaling, indicate a small but consistent effect.

4.4 Discussion

The motivation behind this experiment was to investigate the extent to which listeners can learn to distinguish and identify voices that vary in average f0 and FF-scaling. Results indicate that listeners are able to report voice FF-scaling with reasonable accuracy after only a short training session. Performance was much higher than chance in both the Two-Factor Task and the FF-scaling only Task, for absolute identifications of voice FF-scaling and f0 where applicable. The high rate at which listeners are able to absolutely identify voice FF-scaling is noteworthy given that the ΔFF-scales used in this experiment (7-10%) are not much higher than the just noticeable difference in FF-scaling, which has been estimated to be between 4-8% (Smith et al. 2005, Ives et al. 2005). Furthermore, listeners are able to report voice FF-scaling with the same level of accuracy whether they are asked to report voice f0 or to disregard it.

In addition to the high rate at which listeners correctly identified stimulus FF-scaling, their errors tended to cluster around the correct stimulus FF-scaling. Overall, in 65% of errors committed across both tasks, listeners were only off by a single FF-scaling step. In the Two-factors Task, listeners erred in identifying

stimulus FF-scaling by a single step in 39.7% of trials. Combined with the 40.1% of cases in which they correctly identified voice FF-scaling, this means that in 79.8% of trials listeners were either correct or off by a single step. In the FF-scaling only Task, they were within one FF-scaling step in 78.8% of cases. By chance alone, listeners would be expected to respond within one step of correct in 52% of cases, meaning that they responded within one step roughly 53% (i.e. (79-52)/52) more than expected. These near-miss error patterns suggest that the listener-internal mappings of the stimulus voices are arrayed in a two-dimensional space corresponding closely to f0 and FF-scaling. These results all support the notion that there exists a perceptual quality, such as pFF-scaling, which is closely aligned with FF-scaling.

The ability listeners have demonstrated in reporting voice FF-scaling suggests that the experiment reported here could easily be extended to investigate the relationship between apparent speaker gender and pFF-scaling by instructing listeners that the speaker was of a particular gender on a given trial. A methodology of this kind could be used to investigate the results presented in Johnson et al. (1999) and Glidden and Assmann (2004) where changing listener expectations regarding speaker gender affected perceived vowel quality. If trained listeners systematically over or underestimated stimulus FF-scaling based on apparent speaker gender, it would serve as good evidence that apparent speaker gender affects perceived vowel quality by affecting pFF-scaling estimates based on gender stereotypes.

Another possibility is the use of this training experiment in conjunction with experiments such as those described in Johnson (1990), Johnson et al. (1999) and Barreda and Nearey (2012a), in which the relationship between apparent speaker characteristics and perceived vowel quality was investigated. In those experiments, stimulus vowels varied along a limited number of FF dimensions (either F1 or F1 and F2) rather than along all FFs simultaneously, which is the case when they vary in terms of FF-scaling. For example, in Barreda and Nearey (2012a) listeners were presented with vowels that varied along an F1-F2 continuum, and these were presented with several different f0 and higher formant

conditions. Apparent speaker size and gender judgments were collected in order to control for estimates of pFF-scaling, and the association between these characteristics and perceived vowel quality was investigated.

However, the results of the experiment presented here suggest that it is possible to ask trained listeners to report speaker FF-scaling directly. For example, given a certain point along the F1-F2 continuum, we might expect that listeners would respond to changes in the higher formants by indicating different judged FF-scaling levels. Furthermore, given a point along the F1-F2 continuum, pFF-scaling may co-vary with apparent speaker gender, and perceived vowel quality. Using a methodology of this kind, the relationship between pFF-scaling, apparent speaker characteristics and perceived vowel quality could be investigated more directly.

Barreda and Nearey (2012b) present preliminary results of a study using just this methodology. A replication of Barreda and Nearey (2012a) was carried out in which FF-scaling judgments, as well as speaker gender and vowel quality judgments, were collected from trained listeners. The results indicated that a significant relationship between listener FF-scaling responses and reported vowel quality for vowels which had been low-pass filtered above F3[19].

Although listeners are able to report stimulus FF-scaling accurately, some results suggest that the determination of pFF-scaling interacts with the identification of stimulus f0 in a complicated manner that warrants further investigation. Correct identification of stimulus f0 was associated with higher correct identification of FF-scaling both between-participants (as reported in Section IIIA1) and within participants: of the 46 listeners who made at least five f0 identification errors, FF-scaling identification rates were 6.3% higher when they identified f0 correctly relative to cases in which they did not [$t(45) = 2.91$, $p = 0.0056$].

[19] However, to our surprise, this was not the case for vowels with more higher-formants. The presence or absence of higher formants had a complicated relationship with apparent speaker gender and reported pFF-scaling. This may have resulted in a weakening of the relationship between reported pFF-scaling and reported vowel quality.

Furthermore, a significant negative correlation was found in errors of f0 and FF-scaling identification. A number may be assigned to judged f0 and FF-scaling that indicates the difference between these judgements and the veridical stimulus properties. So, for example, zero would indicate a correct identification while negative integers would indicate underestimations and positive numbers would indicate overestimations. For the 46 listeners who made at least 5 f0 identification errors, the average within-participant Spearman's correlation coefficient between f0 and FF-scaling identification errors was -.17 [$t(45) = -5.88$, $p < 0.0001$] indicating that FF-scaling overestimations were associated with f0 underestimations and vice versa.

The results presented in Section IIIB indicate that stimulus f0 has a weak effect on judged FF-scaling, and that this effect can vary for particular combinations of f0 and FF-scaling. Furthermore these relationships may vary based on the specific task at hand. For example, in the Two-factors Task, there was no significant main effect for stimulus f0 on judged FF-scaling, while for the FF-scaling only Task the main effect for stimulus f0 was significant. This may indicate that f0 has more of an effect on judged FF-scaling when listeners do not have to explicitly report it, relative to situations in which they do have to report it.

The main effect of f0 on judged FF-scaling was positive in cases where it was significant. This is not surprising given the natural co-variation of f0 and FF-scaling, where higher f0s are associated with higher FF-scalings, and the fact that listeners have demonstrated a sensitivity to this covariation (Assmann and Nearey 2007, Assmann and Nearey 2008). However, we do not have ready explanations for the interaction patterns observed across the two tasks. In both cases, the linear relationship between stimulus and judged FF-scaling differs for the middle f0 step relative to the high and low f0 steps, and this difference manifested itself as a decrease in the positive relationship between the two variables, resulting in a compression towards the middle of the response space.

These results suggest that f0 may play a role in the determination of pFF-scaling, and that this may not be determined solely on the basis of the FFs of a vowel sound. An effect for f0 on pFF-scaling is predicted by Method 6 of the

sliding template model of Nearey and Assmann (2007), where they suggest that pFF-scaling (which they refer to as Ψ) is determined party on the basis of f0. However, this model would only predict linear shifts in pFF-scaling based on f0, and not a complicated pattern of interactions. This model also has no way to explain the negative correlation of errors observed, nor the varying effect of f0 based on task type.

The significant and complicated effect of stimulus f0 on judged FF-scaling casts doubt on the theories put forth by Irino and Patterson (2002), Smith et al. (2005), and Turner et al. (2006). These researchers claim that the peripheral auditory system performs transformations on speech sounds that automatically segregate information related to vocal-tract configuration from information related to FF-scaling, and that human listeners have direct access to FF-scaling information resulting from this processing. If this were the case, there is no clear reason why f0 should significantly influence directly reported FF-scaling judgements, or for this influence to vary based on task. Although transforms such as those suggested by these authors may still occur, a transformation which segregates information regarding voice FF-scaling, only to recombine it with f0 information before the listener can access it would not be of much use to listeners.

Some characteristics of the experimental design selected with that goal in mind make it unsuitable to answer detailed questions regarding the processes that underlie the construction of a pFF-scaling dimension, and the manner in which this is influenced by f0. This experiment was designed to investigate whether listeners are able to identify voices on the basis of their FF-scaling, and whether it would be feasible to collect FF-scaling estimates from listeners in perceptual experiments.

First, the sampling of the f0 dimension was deliberately sparse, and many listeners committed very few, or no f0 identification errors at all. For example, 35 of 71 listeners made less than 5 f0 identification errors out of a total of 45 trials for the Two-factors Task. We did not want to present too complex or frustrating a task to listeners until we were certain they could reliably respond to FF-scaling differences in voices. Secondly, the sampling of the FF-scaling dimension was

intended to replicate the stimulus design of experiments that might involve the collection of FF-scaling estimates rather than to investigate the process of FF-scaling estimation as a continuous dimension. Finally, the limited number of trials carried out for each of the two tasks makes it difficult to analyze these processes in great detail. However, it is important to note that the effect of f0 and the correlation of errors were detectable despite these shortcomings, which suggests that these are important considerations in the construction of a pFF-scaling dimension.

In the future, experiments with stimuli that more densely sample the f0 x FF-scaling space, and which feature a higher number of trials will need to be carried out to investigate more specific questions regarding the processes involved in f0 and FF-scaling estimation. Of particular interest to the field of speech perception is the way in which these two processes may cooperate and the ways in which this cooperation may interact with the estimation of apparent speaker characteristics and the determination of vowel quality.

4.5 Conclusion

The experiment outlined here involved a training method in which listeners learned to report voice FF-scaling. Although listeners have previously demonstrated a sensitivity to changes in voice FF-scaling independently of f0, the average listener may not have a ready label for the acoustic characteristic associated with the average FFs produced by a voice. Results indicate that listeners are able to provide FF-scaling judgments with relative ease and consistency, and that these estimates are most strongly determined by the FFs of a stimulus, with only weak effects for stimulus f0. This may be contrasted with apparent speaker characteristics such as apparent speaker size and gender, which are most strongly determined by the f0 of a vowel, with a weaker effect for the FFs (Gelfer and Mikos, 2005; Hillenbrand and Clark, 2009).

The results presented here suggest that it is feasible to collect FF-scaling estimates from listeners in further experiments which seek to investigate the process of FF-scaling estimation, or the role of FF-scaling estimation in speech perception. Furthermore, they suggest that there exists a perceptual dimension

closely aligned with FF-scaling (i.e., pFF-scaling), and that this perceptual dimension may be influenced to some extent by f0 in a complicated manner that is not explained by any theory we are aware of. Given the potential importance of FF-scaling, and its perceptual counterpart pFF-scaling, for vowel perception and the determination of apparent speaker characteristics, these issues warrant further investigation.

Works Cited

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant and M. Tatham (Eds.). Auditory analysis and perception of speech. London: Academic Press. 103-113.

Assmann, P.F. and Katz, W.F. (2000). Time-varying spectral change in the vowels of children and adults. J. Acoust. Soc. Am. 108(4): 1856-1866.Assmann P.F., Nearey T.M. (2008). Identification of frequency-shifted vowels. *Journal of the Acoustical Society of America* 124(5), 3203-3212.

Assmann P.F. and Nearey T.M. (2007). Relationship between fundamental and formant frequencies in voice preference. Journal of the Acoustical Society of America 122(2), EL35-EL43.

Barreda, S. and Nearey, T. (2012b). The association between speaker-dependent formant space estimates and perceived vowel quality. Canadian Acoustics 40: 12-13.

Barreda, S. and T.M. Nearey. (2012a). The direct and indirect roles of fundamental frequency in vowel perception. Journal of the Acoustical Society of America 131: 466-477.

Bland, J. M., & Altman, D. G. (2011). Correlation in restricted ranges of data. *BMJ*, *342*.

Collins, S. A. (2000). Men's voices and women's choices. Animal Behavior 60: 773–780.

Fant, G. (1960). Acoustic Theory of Speech Production. The Hague: Mouton. pp.107-138.

Fant, G. (1975) Non-uniform vowel normalization, STL-QPSR 2-3: 1 – 19.

Fitch, W. T. (1994). Vocal Tract Length Perception and the Evolution of Language. Doctoral dissertation, Brown University.

Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging Journal of the Acoustical Society of America 106: 1511-1522.

Fujisaki, H. and Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels. IEEE Transactions on Audio and Electroacoustics AU-16, 73-77.

Gelfer, M.P., & Mikos, V.A. (2005). The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based on Isolated Vowels, *Journal of Voice* 19: 544-554.

Gonzalez, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics* 32: 277-287.

Glidden, C. and Assmann, P. F. (2004). Effects of visual gender and frequency shifts on vowel category judgments. Acoustics Research Letters Online 5: 132-138.

Gumpertz, M., and Pantula, S. G. (1989). A Simple Approach to Inference in Random Coefficient Models. The American Statistician, 43(4), 203-210. doi:10.2307/2685362

Hollien, H., Green, R., and Massey, K. (1994). Longitudinal research on adolescent voice change in males. Journal of the Acoustical Society of America 96: 2646–2653.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America*, *117*(4), 2193–2200.

Hillenbrand, J.M., Getty, L.A., Clark, M.J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America, 97, 3099-3111.

Hillenbrand, J.M., and Clark, M.J. (2009). The role of F0 and formant frequencies in distinguishing the voices of men and women. Attention, Perception, and Psychophysics, 71, 1150-1166.

Irino, T., and Patterson R. D. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. *Speech Commun*. 36, 181–203.

Ives, D. T., Smith, D. R. R. and R. D. Patterson. (2005). Discrimination of speaker size from syllable phrases. Journal of the Acoustical Society of America 118: 3816-3822.

Johnson, K. (1990). The role of perceived speaker identity in f0 normalization of vowels. Journal of the Acoustical Society of America 88: 642 – 654.

Johnson, K., Strand, E. A. and D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. Journal of Phonetics 27: 359-384.

Johnson, K. (2005). Speaker Normalization in speech perception. In Pisoni, D.B. and Remez, R. (eds) The Handbook of Speech Perception. Oxford: Blackwell Publishers. pp. 363-389.

Joos, M. (1948). Acoustic phonetics. Language 24, 1–136

Katz, W.F. and Assmann, P.F. (2001). Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing. *Journal of Phonetics* 29, 23-51.

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America 29: 98–104.

Lass, N. J., and Brown, W. S. (1978). Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. Journal of the Acoustical Society of America 63: 1218–1220.

Lee, S., Potamianos, S. and Shrikanth Narayanan. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. Journal of the Acoustical Society of America 105: 1455-1468.

Nearey, T. M. (1978). Phonetic Feature Systems for Vowels. PhD thesis, Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America 85. 2088–2113.

Nearey, T. M. and Assmann, P. F. (2007). Probabilistic "sliding template" models for indirect vowel normalization. In Maria-Josep Solé, Patrice Beddor, and Manjari Ohala (eds.) Experimental Approaches to Phonology. Oxford: Oxford University Press. 246-69.

Miller, R.L. (1953) Auditory tests with synthetic vowels. Journal of the Acoustical Society of America 25, 114-121.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America 24: 175-184.

Rendall, D., Vokey, J. R., and Nemeth, C.. (2007). Lifting the Curtain on the Wizard of Oz: Biased Voice-Based Impressions of Speaker Size. Journal of Experimental Psychology: Human Perception and Performance 33: 1208 –1219.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: an expanded typology. *Journal of Applied Psychology*, *85*(1), 112.

Smith, D. R. R. and R. D. Patterson. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. Journal of the Acoustical Society of America 118: 3177-3186.

Smith, D. R. R., Patterson, R. D., Turner, R, Kawahara, H. and T. Irino. (2005). The processing and perception of size information in speech sounds. Journal of the Acoustical Society of America 117: 305-318.

van Dommelen, W. A. and Moxness, B. H. (1995). Acoustic Parameters in Speaker Height and Weight Identification: Sex-Specific Behaviour. Language and Speech 38: 267-287.

Turner, R. E., Al-Hames, M. A., Smith, D. R. R., Kawahara, H., Irino, T., and Patterson, R. D. (2006). Vowel normalisation: Time-domain processing of the internal dynamics of speech. in Dynamics of Speech Production and Perception, edited by P. Divenyi. Amsterdam: IOS Press. pp. 153-170.

Chapter 5

Discussion

In this chapter I will summarize the results of the experiments presented in this thesis, and outline a refined theory of active speaker normalization based on these results. The proposed normalization procedure was implemented as a computer algorithm based on the Sliding Template Model of Nearey and Assmann (2007) but extended to simulate active cognitive control organized around the probabilistic detection of speaker changes. This new model was applied to simulate the results of Experiments 1 and 2. The simulated results will be compared to those observed for human listeners, and will be contrasted with the behaviour of alternative theories of vowel normalization.

5.1 Summary of results

5.1.1 Experiment 1

In Experiment 1, listeners were presented with a seven-step F1-F2 vowel continuum. Each step along this continuum was matched with three f0 and three F3+ (F3 and higher formants) levels, resulting in 63 unique vowel sounds. Listeners were presented with these vowels one at a time in a fully-randomized condition, blocked by repetition. For each trial, listeners were asked to report the category of the vowel (either /ʌ/ or /æ/) and the gender and size of the apparent speaker.

Two analyses of the results were presented. In the first, a partial-correlation analysis was undertaken which found the strength of the independent linear relationship between all combinations of pairs of stimulus variables (F1-F2, f0, F3+) and response variables (speaker size, speaker gender, vowel openness). This analysis revealed that apparent speaker gender is significantly related to vowel openness, even after controlling for the acoustic characteristics of the

stimuli. This result is potentially problematic for cognitively-passive pure-intrinsic theories of vowel perception. For example, Syrdal and Gopal (1986) and Patterson and colleagues (Irino & Patterson, 2002; Smith et al., 2005) have both proposed theories that explain perceived vowel quality solely in terms of the acoustic characteristics of individual vowel tokens. Based on these theories, there is no reason to expect that any apparent speaker characteristic would be significantly related to perceived vowel quality independently of the acoustic characteristics of the sound.

On the other hand, the significant correlation of gender and vowel openness can be accommodated by theories of speaker normalization. In certain cases, shifts in the reference space may result in changes in perceived vowel quality. For ambiguous vowels, such as those used in Experiment 1, differing FF-scaling estimates may result in shifts in vowel quality. Since apparent speaker gender is strongly determined by voice FF-scaling (Hillenbrand & Clark, 2009), we would expect that apparent speaker gender will provide us with information about the reference space being used by a listener. As a result, in some cases, apparent speaker characteristics such as speaker gender may be significantly related to perceived vowel quality.

The second analysis involved comparing the partial correlations between certain pairs of variables in two situations: when apparent speaker characteristics were controlled for, and when these characteristics were not controlled for. As outlined in Section 2.4, a purely indirect effect should only affect perceived vowel quality in situations where it affects the reference space. Consequently, the difference in the strength of partial correlations between the two aforementioned conditions can serve as a rough index of the degree of indirectness of effects on vowel quality. This analysis revealed that the partial correlation between f0 and vowel openness decreased by 63% on average, while those between F1 and F3+ and vowel openness decreased by only 2.6% and 7.5% respectively. These findings largely support the assertion made by Johnson (1990) that the effect of f0 is primarily indirect, in addition to supporting the suggestion made by Nearey

(1989) that F3 may have some indirect effect by providing listeners with information regarding a speaker's formant-space.

5.1.2 Experiment 2

In Experiment 2, listeners were presented with vowels produced by a series of voices, and were asked to monitor for a single vowel category. They were asked to respond as soon as they heard this vowel category, and to ignore all others. There were six synthetic stimulus voices. These differed along 3 FF-scaling steps (referred to as formant-space steps in Chapter 3) and 2 f0 levels. Vowels were presented in blocks where all vowels were produced by either a single voice, or two different voices.

This design was intended to investigate the contextual tuning theory of Nusbaum and Magnuson (1992) and active speaker normalization in general. According to contextual tuning theory, listeners are expected to refine their reference space until a stable mapping is achieved. If a speaker change is detected, the current frame of reference may be discarded, and a new representation is established. The detection of speaker changes and the refinement of the speaker representation are cognitively-active processes and are expected to be associated with increased reaction times.

In Experiment 2, listeners most accurately identified target vowels when voices had the same FF-scaling and the same f0 level (in a single-voice condition). In the absence of f0 differences between voices, larger FF-scaling differences between voices (i.e., larger reference space mismatches), led to progressively worse identification performance among listeners. However, when FF-scaling differences were accompanied by f0 differences, the negative effect associated with FF-scaling differences between voices was significantly diminished. As presented in Appendix 2, f0 differences between stimulus voices in a block were very likely to result in multiple perceived voices regardless of FF-scaling differences, while FF-scaling differences alone were unlikely to result in the perception of multiple voices.

Essentially, FF-scaling differences between voices in a block led to larger decreases in performance exactly in cases where listeners were unlikely detect

changes in speaker. This suggests that the negative effect associated with mixed-speaker listening conditions may be due, at least in part, to difficulties associated with the detection of speaker changes and the use of inappropriate extrinsic information that may not be appropriate for the current speaker.

In cases where speaker changes are detected, processes related to the recalibration of the reference space are expected to result in increases in response times. However, since these increased response times are related to the more accurate identification of vowel sounds, we expect that in cases where listeners take longer because they are carrying out processes related to normalization reaction times will be positively correlated with identification rates.

There was a significant (negative) marginal correlation between reaction times and identification accuracy meaning that, overall, listeners took longer to respond in blocks where they also responded less accurately. However, in blocks where voices had different f0 levels and speakers were likely to detect speaker changes, listeners took longer to respond, but did not exhibit the decrease in accuracy that might be expected given the negative marginal correlation between response times and accuracy. An interpretation of this result is that, in general, listeners take longer to respond in blocks where vowels are generally difficult to identify, explaining the negative marginal correlation between accuracy and response times. In contrast, in the presence of detected speaker changes, listeners are carrying out processes related to speaker normalization that result in increases in accuracy but also come with a cognitive-cost. Consequently, in cases where listeners are likely to detect speaker changes a positive relationship between accuracy and response times is expected.

Finally, evidence was presented that in the absence of detected speaker changes, listener accuracy improved in single-voice blocks, supporting the notion that listeners refine their FF-scaling estimates throughout a listening situation in the absence of detected speaker changes. When speaker changes were likely to be detected, hit rates were stable within blocks.

5.1.3 Experiment 3

Experiment 3 was intended to investigate whether listeners could learn to report the FF-scaling of voices directly, rather than indirectly by reporting apparent speaker characteristics that are correlated with FF-scaling. Participants took part in a training game where they learned to identify voice FF-scaling using 15 unique stimulus voices (5 FF-scalings levels crossed with 3 f0 levels) where each voice was associated with a different response button arranged on a board. Listeners were played a voice and were asked to click on the button associated with the voice, thereby providing an f0 and FF-scaling estimate for the voice they had just heard. After training, listeners took part in two testing levels intended to assess their ability to report FF-scaling. In the first testing level, listeners were asked to identify both the f0 and FF-scaling level of stimulus voices. In the second testing level, listeners were asked to report only FF-scaling, and ignore stimulus f0.

Results indicate that listeners are able to report voice FF-scaling with a good degree of accuracy and consistency after only a short training session. There was no significant difference in this ability when listeners were asked to report f0, compared to when they were asked to disregard it. Furthermore, error patterns suggest that in cases where listeners did make FF-scaling identification errors, these tended to be clustered around correct FF-scaling levels, suggesting that listeners were in fact responding to an acoustic dimension correlated with FF-scaling. Finally, results suggest a complicated relationship between perceived f0 and perceived FF-scaling. Stimulus f0 level was found to significantly influence FF-scaling estimates, and there was some indication that f0 and FF-scaling errors are negatively correlated[20].

These results show that although f0 can affect perceived vowel quality by affecting FF-scaling estimates, listeners can deprioritize f0 information when making FF-scaling estimates in situations where this is known to provide

[20] The negative correlation of errors was investigated further in Barreda and Nearey (To Appear). The results of that experiment indicate that this correlation is significant and consistent.

unreliable information regarding the FF-scaling of a voice. For example, it was argued that the results of Experiment 1 demonstrate that f0 affects vowel quality mainly by influencing listener-internal FF-scaling estimates, which in turn affect the location of the reference-space. If the same behaviour were seen here, listener FF-scaling responses would be expected to be strongly influenced by stimulus f0, rather than only the weak effect observed in Experiment 3.

5.2 An explicit model of Active Speaker Normalization

The results of these experiments suggest that vowel identification is carried out by a process of speaker normalization, which is governed by a cognitively-active control structure broadly similar in character to the contextual tuning theory of Nusbaum and Morin (1992). The process may be summarized as follows. When a listener encounters speech from a new speaker, the listener uses the intrinsic properties of that speech to estimate an appropriate FF-scaling for the speaker. Because this initial estimate is informed solely by the intrinsic properties of the vowel to be classified, this will be referred to as the *intrinsic* FF-scaling estimate. This parameter will determine the reference space used to interpret vowels produced by that speaker. Cues may affect vowel quality indirectly, by affecting the estimation of this parameter. For example, telling the listener that a speaker is female, or playing a vowel with a high f0, may both result in the expectation of a relatively high FF-scaling, which may then affect perceived vowel quality.

The intrinsically-specified FF-scaling estimate will then become the basis of the running FF-scaling estimate for that speaker going forward. This estimate may be updated based on new information regarding the appropriate FF-scaling for the speaker. Because this estimate potentially includes information extrinsic to the vowel to be classified, this will be referred to as the *extrinsic* FF-scaling estimate.

The categorization of following sounds then depends on the detection, or lack thereof, of a change in speaker. In the event that a speaker change is detected, an FF-scaling estimate can be calculated based solely on the intrinsic properties of the current vowel sound. This intrinsic FF-scaling estimate may then replace the

current extrinsic estimate, reflecting the fact that previous information is no longer useful, and indirect cues may strongly influence perceived vowel quality by influencing the new estimate.

If a speaker change is not detected, one of two things may occur. If the extrinsic FF-scaling estimate is deemed to be unstable and not exactly appropriate, the estimate may be updated based on the properties of the current stimulus. In the event that a speaker change is not detected and the current extrinsic FF-scaling estimate is deemed to be stable, the vowel stimulus is classified using the existing FF-scaling estimate, eliminating the additional processing associated with the estimation or refinement of the running FF-scaling estimate. In these cases, indirect cues (i.e., cues that affect vowel quality only by affecting FF-scaling) should lose their effect on perceived vowel quality.

In the remainder of this section, Method 6 of the Sliding Template Model will be described. Following this, an explicit model of active speaker normalization will be outlined. This model provides something like an active control structure for Nearey and Assmann's (2007) Method 6, based on the results of the experiments summarized above. Because this model is meant to replicate a normalization method with active-cognitive control over certain processes, it will be referred to as the Active Sliding Template Model (ASTM).

5.2.1 The Sliding Template Model

The Sliding Template models of Nearey and Assmann (2007) were designed to account for identification of vowels in a mixed speaker condition, where both vowel category and speaker identity vary randomly from trial to trial. These models predict perceived vowel quality by estimating an appropriate FF-scaling for a vowel sound and modifying the observed formant-pattern to compare it to the reference patterns specifying expected FFs for different vowel categories. The vowel category whose reference pattern provides the closest fit to the observed formant-pattern, given the estimated FF-scaling, is selected as the winning vowel category.

The authors describe several models that differ in the ways they estimate FF-scaling or in the manner that they specify the reference vowel-patterns. Of

particular interest is their preferred method, Method 6 of the Sliding Template Model, which will be outlined here. Throughout this discussion and in keeping with Nearey and Assmann, FF-scaling will be indexed using the log-mean F1-F2-F3 frequency across a speaker's entire vowel inventory. This measure provides an adequate way to compare the reference spaces of different speakers of the same dialect under the assumption of uniform scaling within vowel category, between speakers.

Selecting an FF-scaling for a candidate vowel category that maximizes the fit between observed and expected FFs without considering the distributional properties of f0 and FF-scalings could lead to improper FF-scaling estimates. For example, a winning FF-scaling appropriate for a small child could be predicted for a vowel with an f0 of 100 Hz, an extremely unlikely pairing of acoustic characteristics in the real world. Furthermore, the winning FF-scaling could be well outside the normal range of FF-scalings, potentially leading to a high rate of misclassifications. Method 6 attempts to remedy these issues by selecting an FF-scaling estimate for each candidate vowel category that maximizes the fit between the observed formant-pattern and that expected for each vowel category, while also taking into account the (approximate) joint distribution of f0 and FF-scaling across a human population.

The fit between observed and expected formant patterns given an FF-scaling estimate is quantified with reference to a multivariate normal distribution where the mean vector corresponds to the formant reference-pattern for that vowel category and the covariance matrix is the pooled within-category covariance matrix provided to the model. Since the goodness of the fit between observed and expected FFs varies according to a single parameter (FF-scaling), the probability density function associated with this consideration is univariate normal. The mode of the density of this distribution alone will correspond to the FF-scaling which results in the best fit between observed and expected formant patterns.

However, as mentioned above, the FF-scaling that leads to the best fit could be implausible given the distributional properties of f0 and FF-scaling. For this reason, the aforementioned probability density function is multiplied by the

conditional probability of f0 given FF-scaling, and the prior probability of FF-scaling. The FF-scaling that maximizes these considerations can be found analytically by finding the product of these densities, and finding the mode of the resulting density. The mode of this density corresponds to the winning FF-scaling estimate for the vowel category.

The conditional probability of f0 given FF-scaling is determined based on the linear relationship between log f0 and FF-scaling across a range of speakers, using the same parameter values suggested in Nearey and Assmann (2007). This relationship is presented visually in Figure 5.1. The expected f0 given FF-scaling was found using equation (1), where FF-scaling is represented by 'FFS' to avoid ambiguity:

$$(1)\quad \widehat{f0} = 2.14452 \bullet (FFS) - 10.3233$$

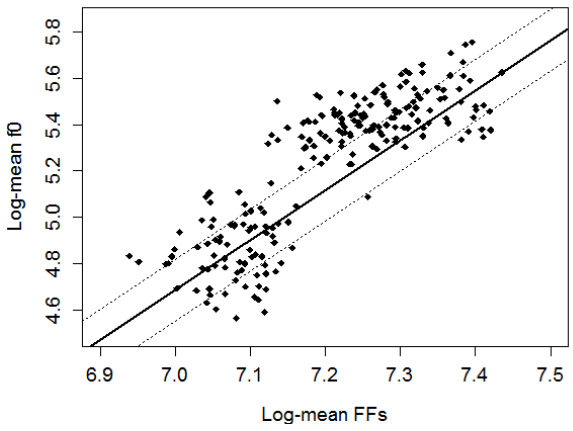

Figure 5.1. *Scatterplot of speakers from two data sets (Peterson and Barney 1952; Hillenbrand et al. 1995) plotted according to their FF-scaling (indexed by log-mean FFs) and log-mean f0. The bold line indicates the regression line predicting log f0 on the basis of FF-scaling. The dotted lines parallel to the regression line indicate one standard deviation in f0 given the FF-scaling.*

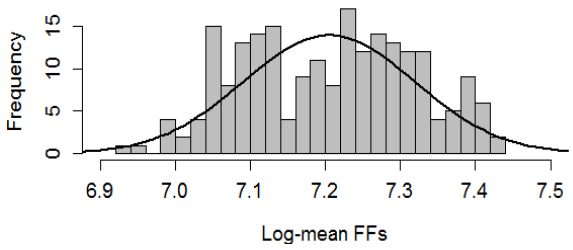

Figure 5.2. *Histogram of the marginal (prior) distribution of FF-scalings from across two data sets (Peterson and Barney, 1952; Hillenbrand et al. 1995). The bold line shows the density of a normal distribution with the same mean and variance parameters as this marginal distribution.*

Based on the linear relationship presented in (1), an f0 may be predicted for candidate FF-scalings, and the distance between the observed f0, and the f0 predicted for the FF-scaling may be found. This distance is then penalized with reference to the standard error of the regression model. This may be thought of as centering a normal distribution at the f0 predicted for each FF-scaling with a standard deviation equal to the standard error of the regression model, and then finding the probability of drawing the observed f0 from this distribution. The result of this is that FF-scaling estimates that predict an f0 that is close to the observed f0 are prioritized over those that are not. The mode of the conditional distribution of f0 given FF-scaling will be located at the FF-scaling that predicts the observed f0.

The standard deviation of the conditional distribution of f0 given FF-scaling was set to 0.09382. This value is smaller than that suggested in Nearey and Assmann by a factor of 0.707, meaning that the effect for f0 will be relatively stronger. A justification for this is given in in Section 5.3.1.1. The mean and standard deviation of the marginal probability of observing an FF-scaling were set to $\mu = 7.2333$ and $\sigma = 0.1284$, based on the values suggested by Nearey and Assmann.

Once an optimal FF-scaling estimate has been found for each vowel category, the winning vowel category is the one that provides the best match for the observation, given its category-specific FF-scaling. This may be determined by subtracting the category-specific FF-scaling estimate from the observed formant-pattern, and finding the minimum Mahalanobis distance[21] between the observed formant-pattern and that expected for the vowel category. Nearey and Assmann (2007) describe the process of selecting the best fitting vowel category, given the best possible FF-scaling for that category as "choose the vowel that looks best when it tries to look its best" (p. 235).

[21] Mahalanobis distances are multivariate measurements of distance that take the covariance patterns of variables into account. Unless otherwise specified, all references made to the calculation or comparison of distances refers to Mahalanobis distances calculated using the covariance matrix presented in Table 5.2.

Vowel	F1	F2	F3
i	-1.4191406	0.62950438	0.8487829
ɪ	-0.9816006	0.41787938	0.7149399
e	-0.9211756	0.49334438	0.7382749
ɛ	-0.7035106	0.32145438	0.7001354
æ	-0.4457406	0.22904938	0.6735179
ɑ	-0.5450806	-0.09682062	0.6557979
ʌ	-0.6572806	0.01781438	0.6558074
o	-0.8547056	-0.17574562	0.6554719
ʊ	-0.9187456	-0.09007562	0.6727639
u	-1.2477756	0.01196438	0.6208954

Table 5.1. *Reference patterns specifying the expected F1 F2 and F3 frequencies (in normalized log-Hz) for the vowel phonemes of Edmonton English. If an FF-scaling is added to these values and the sum is exponentiated, the FFs (in Hz) expected for each vowel category given the FF-scaling estimate may be found.*

	F1	F2	F3
F1	0.0147141	0.0010423	-0.0010114
F2	0.0010423	0.0100548	0.0000059
F3	-0.0010114	0.0000059	0.0056742

Table 5.2. *Pooled within-groups covariance matrix given to the ASTM to be used for the classification of vowel sounds.*

The reference patterns used for each vowel category, as well as the pooled within-groups covariance matrix used to classify vowels are given in Tables 5.1 and 5.2 respectively. By convention, the sum of the reference patterns specifying expected FFs across all vowel categories is equal to zero. This may be achieved by specifying the reference patterns using formant values normalized using the log-mean normalization method of Nearey (1978). The reference patterns were determined relative to vowel data collected from Edmonton English speakers (Thomson, 2007), while an appropriate pooled within-groups covariance matrix

was estimated using data collected from a large data set (Peterson and Barney, 1952).

5.2.1.1 Control structure implied by the Sliding Template Model

Nearey and Assmann indicate that the Sliding Template Model could be modified to accommodate different uses of f0 or prior information in estimating FF-scaling. This can be carried out by changing the parameters specifying the probability distributions that are used to determine the winning FF-scaling estimates. For example, the relative strength of f0 can be modified by increasing or decreasing the variance of the conditional distribution of f0 given FF-scaling, where decreasing this variance results in a stronger effect for f0. Furthermore, the influence of a priori information regarding FF-scaling can be manipulated by changing the mean or variance of the marginal distribution of FF-scaling.

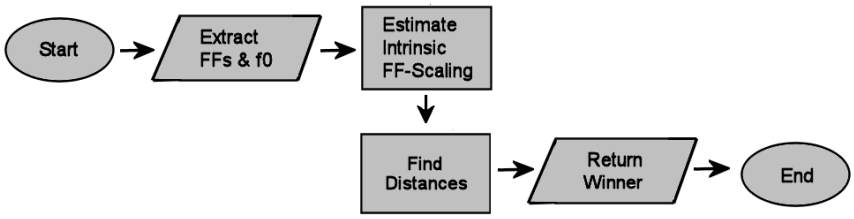

Figure 5.3. *A flowchart representing the control structure implied by the Sliding Template Model as described in Nearey and Assmann (2007).*

However, the original Sliding Template Model is governed by an open-loop control structure, so that there is no mechanism to implement these changes from trial to trial. As seen in Figure 5.3, there is a single path from input to output, and there is no mechanism by which feedback or any decisions related to the detection of speaker changes may affect the outcome of the process. Essentially, Method 6 of the Sliding Template Model is a pure-intrinsic method of FF-scaling estimation with no memory, and a control structure appropriate for a cognitively-passive process.

5.2.2 The Active Sliding Template Model

An overview of the proposed control-structure for the Active Sliding Template Model (ASTM) is presented in Figure 5.4. This proposed control structure has two major differences compared to the original Sliding Template Model[22]. First, the ASTM has a memory that keeps track of an extrinsic FF-scaling estimate, and can both refine and discard this estimate as necessary. Second, the ASTM features processes that monitor for speaker changes and for the appropriateness and stability of the current running FF-scaling estimate. The additional processes featured in the ASTM are all related to these two changes. In the following subsections, I will outline the processes composing the ASTM as depicted in Figure 5.4.

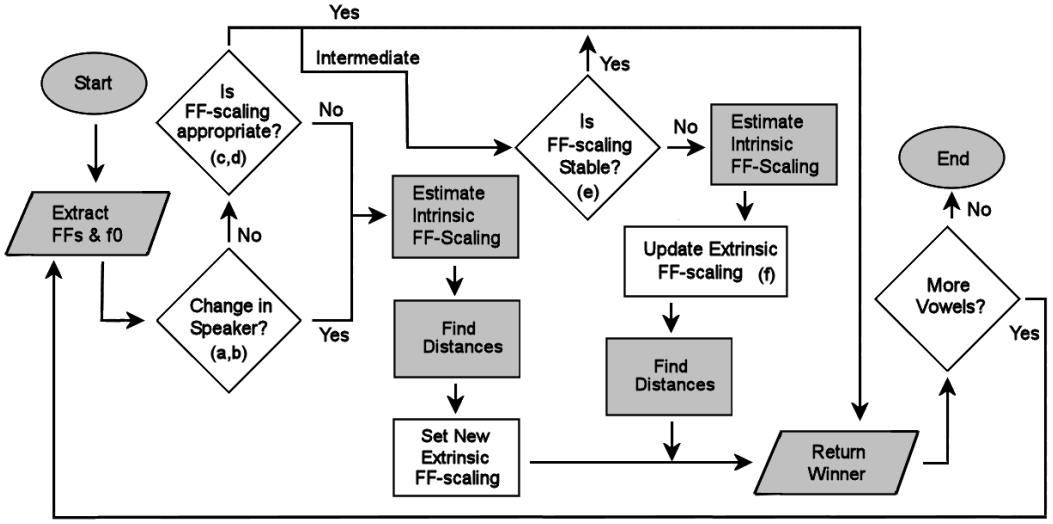

Figure 5.4. *A flowchart representing the processes of the Active Sliding Template Model. The stages shared by the unmodified Sliding Template Model are shaded in grey. The letters in brackets indicate where the parameters outlined in Table 5.3 are used by the model.*

[22] This process has some similarities to Weenink's Speaker-Adaptive normalization method (Weenink, 2006; Ch. 11). That method also features a reference-space that may change from trial to trial to maximize the fit to the current vowel stimulus. However, that model has no role for the detection of speaker changes. Consequently, this model is most similar to a passive-extrinsic model of vowel perception. As discussed in Chapter 3, such models cannot recreate the pattern of results observed in Experiment 2.

The general design of the ASTM followed from the interpretation of the Experiments composing this thesis presented in Section 5.1. No action or state of the ASTM exists solely to replicate specific output patterns without having a theoretical motivation. However, the parameters settings used for the simulation were selected so that they would generate patterns of results like those observed for human listeners. Initial parameter settings were set at levels that were deemed reasonable and successive simulations were run, with parameters being refined in a heuristic manner between repetitions of the simulation. The intent of this was only to investigate whether the model of active speaker normalization outlined above could generate patterns of results similar to that observed for human listeners.

One important difference between the Sliding Template Model and the ASTM is that a small amount of Gaussian error was added to the optimal, category-specific FF-scaling estimates as calculated by Method 6. This was done to reduce the performance of the model, and to create an advantage to refining rather than simply discarding estimates. Error in FF-scaling estimates was implemented by adding Gaussian noise with a standard deviation of 0.083 to the optimal FF-scaling estimate for each category. Since these estimates are expressed in log-Hz, this amounts to an expected error of roughly 8.3% in estimates when considered in linear Hz values. The size of error was determined with reference to observed FF-scaling reporting errors as reported in Barreda and Nearey (To Appear).

5.2.2.1 Summary of tuned parameters.

A total of six tuned parameters were required to implement the Active Sliding Template Model (ASTM). These parameters are presented and summarized in Table 5.3, and will be explained in detail in the remainder of this section. In addition, as outlined above, two changes were made to Method 6 of the Sliding Template Model: the reduced value of the conditional variance of f0 given FF-scaling, and the addition of error to the FF-scaling estimates made by the model to simulate, roughly, perceptual or choice error in listener's assessment of the evidence at hand.

Parameter	Setting	Description
(a) f0-related speaker change detection	0.5	Determines the rate at which the model will detect a speaker change solely based on changes in f0.
(b) f0-related threshold multiplier	1.5	When there has been an f0 change, the minimum distance is multiplied by this value, increasing the apparent lack of fit of the current reference space.
(c) Refinement distance	0.25	Determines the maximal minimum distance that will not lead to refinements of the current FF-scaling estimate.
(d) Speaker change Distance	6.25	Determines the maximal minimum distance that will not lead to a detected speaker change.
(e) Stability parameter	4	If a speaker change is not detected for this many trials, the FF-scaling is assumed to be stable.
(f) FF-scaling refinement combination weight	4	Determines the relative weight of the current FF-scaling estimate when this value is updated.

Table 5.3. *A summary of tuned parameters involved in the Active Sliding Template Model.*

5.2.2.2 Detection of speaker changes

The detection of speaker changes is carried out in two ways. The model is presented with data describing a vowel sound with a given f0 and formant frequencies. If this vowel is the first in a listening situation (e.g., the first in a block or round), a new listener is assumed. If the vowel is not the first in a listening situation, the f0s of the current and previous stimuli are compared, and if these differ, the detection of a speaker change is probabilistically determined[23].

[23] The probability of detecting a speaker change based on differences in f0 would have to be related to the magnitude of this difference to accurately reflect the behaviour of human listeners. However, in the stimulus design of Experiments 1 and 2, the smallest f0 difference is a half octave, and listeners do not have any reason to expect large differences in f0 from trial to trial except where they signal a change in speaker. For this reason, and for the sake of simplicity, differences in f0 were treated as binary (i.e., same vs. different). Simulation of further experiments may require a more nuanced approach to f0-related detection of speaker changes.

This was implemented by drawing a uniform random variable between 0 and 1 and triggering a detected speaker change when this variable was above the f0-related speaker change parameter. Lowering or raising the value of this parameter affects the rate at which listeners will detect a speaker change solely on the basis of f0, regardless of possible FF-scaling differences between the previous and current speaker. By setting this parameter to 1, any f0 change will result in a detected speaker change, while setting this parameter to 0 means that the model will never detect a speaker change on the basis of f0 differences. A parameter setting of 0.5 resulted in a good correspondence of simulation outputs to observed results.

Speaker changes were also signalled based on the appropriateness of the current FF-scaling estimate, the assessment of which is described in the next subsection. If a change in speaker is signaled by any of these mechanisms, FF-scaling is estimated using the information intrinsic to the vowel in the same manner as outlined for Method 6 of the original, non-adaptive Sliding Template Model, save for the addition of estimation error.

5.2.2.3 Assessing the appropriateness of current FF-scaling using the refinement and speaker change distance thresholds

If a speaker change is not detected based on f0 differences or because of a change in listening situation, the appropriateness of the current FF-scaling is tested. This appropriateness was quantified by finding the minimum Mahalanobis distance between the current vowel and the expected locations of each candidate vowel-category given the current reference-space location.

This minimum distance may be interpreted in one of two ways. If the current reference space is assumed to be correct, and produced formant patterns are expected to be probabilistically related to the expected formant patterns for a given speaker, then an increasing minimum distance represents a decreasing probability that the reference pattern associated with the current reference space would generate the observed formant pattern. Alternatively, the reference space could be assumed to be incorrect, and the new vowel could be considered to accurately represent a different underlying FF-scaling. In these cases, the distance

between the observed and expected formant patterns will be directly related to the underlying FF-scaling differences and larger minimum distances would make a single underlying FF-scaling increasingly unlikely.

Two parameters were used to divide minimum distances into three general classes: 'appropriate', 'intermediate' and 'inappropriate'. For reasons which will be made clear in the following paragraphs, the lower threshold, which splits appropriate and intermediate distances, will be referred to as the *refinement distance.* The higher threshold, which splits intermediate and inappropriate distances, will be referred to as the *speaker change distance*.

In cases where the minimum distance was below the refinement distance (a distance of 0.25 was used for this parameter), the current estimate was deemed to be *appropriate*. This lowest threshold was created to allow for some variation in vowel tokens without necessarily concluding that the current mapping is inappropriate. In cases where current FF-scaling was deemed to be appropriate, the vowel was classified using the current extrinsic FF-scaling estimate as is, with no modification. Since the current FF-scaling estimate is used to calculate the minimum distance, the winning vowel in these cases is simply the vowel associated with this distance. This is seen in Figure 5.4 where, in cases where the current FF-scaling estimate is used as is, the model goes directly from the assessment of the appropriateness of the current estimate, to the selection of the winning vowel category. As a result, in cases where a speaker change is not signalled and the current reference-space is a good fit to the incoming formant pattern, the ASTM classifies vowels with no more computations than those normally incurred by the model to monitor for speaker changes.

If the minimum distance was between the refinement and speaker change distances (set at 6.25), the current FF-scaling estimate was deemed to be *intermediate*. An intermediate minimum distance was meant to simulate a situation in which the current FF-scaling was deemed to be a poor fit and a candidate for improvement, however, the fit was not so poor as to warrant a complete renewal of the extrinsic FF-scaling estimate. In cases where the

minimum distance was intermediate, the FF-scaling was possibly refined based on the outcome of an assessment of its stability (discussed in the next subsection).

Finally, in cases where the minimum distance was larger than the speaker change distance, the current FF-scaling estimate was deemed to be inappropriate. This situation was meant to simulate a situation in which the current FF-scaling estimate (and associated reference space) offered such a poor fit to the observed formant pattern that a speaker change was deemed to be likely. In cases where the minimum distance was above the speaker change distance, a speaker change was signaled and an intrinsic FF-scaling estimate was calculated using Method 6 of the Sliding Template Model. When this occurred, the newly calculated intrinsic estimate formed the basis of the new, extrinsic FF-scaling estimate.

In cases where the current and previous vowels had different f0s, the minimum distance was increased by a fixed percentage resulting in an increased sensitivity to FF-scaling mismatches when these came accompanied with f0 differences. This reflected, in effect, 'growing skepticism' about speaker constancy and biased the model towards updating or rejecting the current FF-scaling estimate by increasing the apparent minimum distance relative to the thresholds. This parameter was set at two, meaning that apparent distances effectively doubled in situations where the current vowel sound had a different f0 than the previous vowel sound (but where the f0 difference did not already trigger the probabilistic detection of a speaker difference)

5.2.2.3 Assessing stability of current FF-scaling using the stability parameter

The stability of the current mapping was determined by keeping track of the number of consecutive previous trials in which the minimum distance between the observed formant-pattern and any reference pattern was below the second threshold, and there was no detected speaker change resulting from differences in f0 or changes in block or round. If this number was greater than or equal to the stability parameter, the mapping was deemed to be stable whereas if the number was below the threshold it was not. The stability parameter was set at 4, meaning that after three consecutive trials in which the current mapping was at least somewhat appropriate, it was considered to be stable and no more refinements

were made to the estimate until a speaker change was detected by any mechanism, even in cases where the minimum distance was of intermediate distance. This functionality was put in place to represent the fact that, after a certain amount of experience with a speaker, listeners are expected to stop refining their reference space until a speaker change is detected.

5.2.2.4 Updating Extrinsic FF-scaling Estimate

In cases where no speaker change was detected but the mapping was determined to be unstable, the running extrinsic estimate was updated. This was done by selecting the optimal intrinsic FF-scaling estimate using Method 6 (as outlined above, and including estimation error), and using this to refine the extrinsic FF-scaling estimate. This was done by calculating the weighted mean of the extrinsic and intrinsic estimates where the current extrinsic estimate is given a weight of 4, and the new intrinsic estimate is given a weight of 1. The effect of this is that the extrinsic estimate only moves 20% of the way towards the new estimate, which is meant to simulate a reluctance to dramatically change the FF-scaling estimate in the absence of a detected speaker change[24].

5.2.2.6 Finding distances and selecting the winner

In cases where classification follows a detected speaker change, a category-specific FF-scaling estimate is subtracted from the observed formant pattern to be classified, and this is compared to the reference pattern specifying each category. In cases where the extrinsic FF-scaling estimate has merely been updated, this value is used for every vowel category. In every case, the vowel category associated with the minimum distance to the reference pattern, given the

[24] This stage is similar to the manner in which the reference space moves from trial to trial in Weenink's Speaker-Adaptive Normalization method (2006). In that method, the motion of the current space towards the new space is controlled by the α parameter, which controls the extent of the motion from the current space to the new space. A setting of 0 denotes no change, a setting of 1 denotes a complete replacement, and a setting of 0.5 means the new space will fall exactly in between the current and new locations. An α setting of 0.2 will produce the same effect as the weighted mean used in the ASTM.

FF-scaling used to make the comparison, was selected as the winning vowel category.

As outlined in Section 5.2.2.2, speaker changes were detected, in part, based on the goodness of fit provided by the current FF-scaling estimate to the newly observed formant-pattern. The goodness of fit provided by the current extrinsic FF-scaling was assessed by finding the minimum distance between the candidate reference-patterns and the observed formant-pattern given the FF-scaling. As a result, in cases where the current extrinsic FF-scaling was used without modification, the vowel category associated with the minimum distance to the observed formant-pattern should be selected as the winning vowel category. This meant that in cases where no changes are made to the FF-scaling used for classification, the winning vowel category may be determined with no more computations than those required to monitor for speaker changes.

5.3 Simulation of results using the Active Sliding Template Model

The ASTM was implemented in a computer algorithm using R (R Core Team, 2013), so that it would match the process outlined in Section 5.2.2. This model was then be used to simulate the results of Experiments 1 and 2. Since the focus of Experiment 3 is on the estimation of FF-scaling independently of f0 and detected speaker changes, simulation of results does not present an interesting case for this model.

For Experiment 1, the focus will be on generating the observed shifts in perceived vowel quality associated with the different f0 and higher-formant levels, and on the weakening of the relationship between f0 and vowel quality when apparent speaker characteristics are controlled for. For Experiment 2, the focus will be on generating the observed pattern of hit rates across the different voice-pair types, and on recreating the association between increased processing time and situations in which listeners were likely to detect speaker changes.

5.3.1 Experiment 1

A matrix was created describing the experimental stimuli used in Experiment 1. Three columns contained information specifying the first three FFs,

and a fourth column specified the f0 for that vowel. Each row contained information describing the FFs and f0 for the vowel sound associated with a single experimental trial, across all participants. A fifth column indicated changes in round or changes in participant. The resulting matrix provided the ASTM with enough information to recreate the sequence of stimuli presented to participants.

5.3.1.1 Effects for f0 and the higher formants on vowel quality

In Experiment 1, listeners had to decide whether vowel stimuli sounded more like /ʌ/ or /æ/, and to report the apparent size and gender of the speaker[25]. To reflect these instructions, the model only considered these two vowels as possible candidates. Since the simulation was given information regarding changes in participant and changes in round, the first stimulus for a new participant or round was treated as coming from a new speaker.

The results of Experiment 1 were simulated 10 times to smooth-out the random component involved in the detection of speaker changes and the estimation of intrinsic FF-scaling estimates. Instances where the model returned an /æ/ were coded as 1, while instances of /ʌ/ were coded as 0. The average classification for each stimulus, for each trial, was found. To investigate the effects of f0 and F3+ changes on categorization of vowels, data were pooled across the F1-F2 continuum steps, within F3+ and f0 condition. The results of this analysis are compared with the results of the same analysis performed on the data observed in Experiment 1.

The model had a tendency to over-predict instances of /æ/, by an average of 6.4% overall. However, there was a close correspondence between observed and simulated categorization of individual trials. In 86% of cases, the average response for a trial was of the same category as the observed response for that trial. As seen in Figure 5.3, the ASTM shows the same general trend of effects for f0 and F3+ as seen in the results of Experiment 1, where lower f0 levels and F3+

[25] The ASTM does not explicitly guess the gender and size of the speaker, however, predictions of this kind could be made based on the stimulus properties of the vowels to be classified and the FF-scaling estimated for each trial.

levels are associated with more /æ/ responses overall. Theories of vowel perception that do not include effects for f0 and the higher formants cannot account for such patterns.

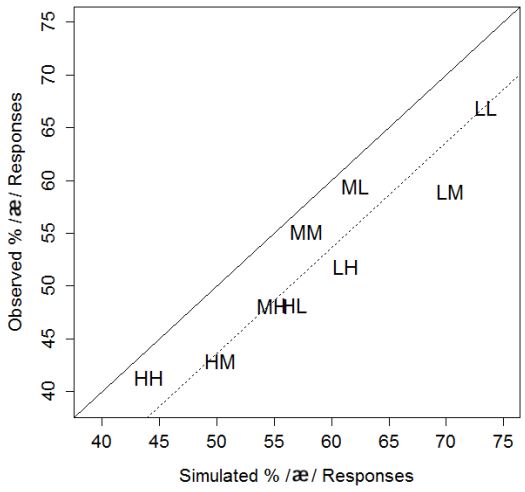

Figure 5.5. *A comparison of observed and simulated percentage of /æ/ responses for the data from Experiment 1. Data is pooled across continuum steps, within F3+ and f0 condition. Letters indicate F3+ (first letter) and f0 (second letter) condition from among: Low (L), Medium (M) and High (H). The solid line indicates points along which x = y. To the extent that the simulation accurately reflects listener behaviour, points should all fall along this line. The dotted line indicates points along which simulated /æ/ responses are 6.4% greater than observed responses.*

As mentioned in Section 5.2.1, the standard deviation of the conditional distribution of f0 given FF-scaling was reduced by a factor of 0.707 relative to the value suggested by Nearey and Assmann (2007). This resulted in a halving of the conditional variance. The decreased standard deviation of the distribution of f0 given FF-scaling was set based on the output of simulations run with the original values. These simulations indicated that the original parameter settings did not result in f0-induced shifts in classification patterns to the extent observed for human listeners in Experiment 1. This was taken as an indication of the fact that the relative strength of f0 information on FF-scaling estimates needed to be increased.

Figure 5.6 compares the results of Experiment 1 to two simulations of these results carried out using the ASTM. These simulations differ solely in terms of the conditional variance of f0 given FF-scaling. It is evident that a variance of half that proposed by Nearey and Assmann led to classification patterns more similar to that observed for human listeners, while the original parameter settings did not show the desired level of sensitivity to changes in f0.

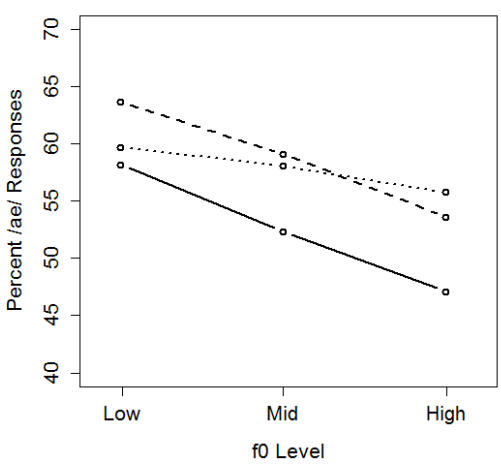

Figure 5.6. *The percent /æ/ responses, organized by f0 level, observed for Experiment 1 are indicated by the solid line. The dotted line shows classification patterns of the ASTM when the conditional variance of f0 given FF-scaling is set equal to the value suggested by Nearey and Assmann (2007). The broken line shows classification patterns by the same model when this parameter is divided in half.*

In simulations reported by Nearey and Assmann (2007), the authors found that they had to increase the conditional variance of f0 given FF-scaling to reflect the behaviour of human listeners, whereas here it was reduced. In the experiment simulated by Nearey and Assmann, listeners were asked to identify speech where, in some cases, there were very large mismatches between FF-scaling and f0 given the normal covariation of these characteristics. As a result, it makes sense that listeners would rely less on f0 to estimate FF-scaling. On the other hand, in Experiment 1, listeners were asked to report the apparent speaker characteristics of unknown speakers, and the phonetic quality of the speech sounds they were asked to identify had no true 'correct' interpretation. In this situation, it makes sense that listeners would rely heavily on f0 to estimate FF-scaling.

The fact that the parameter settings of the ASTM (or the original Sliding Template Model) may need to be modified to accommodate specific listening situations is seen as a strength, rather than a weakness, of the model. It is clear at this point that listeners will adapt their behaviour based on the task at hand, and the information that is deemed useful given the specific listening situations. In light of this, it would be more surprising if a single set of parameter settings were able to accurately reflect the behaviour of human listeners across a range of listening situations. The ASTM, and the framework provided by the Sliding Template Model, features a natural way to accommodate these changing behaviours.

5.3.1.2 Reduction of strength of indirect intrinsic effects

An important result in Experiment 1 consisted of the weakening of the partial correlations between f0 (and to a lesser extent F3+) and vowel openness when apparent speaker characteristics were controlled for. This was taken as an indication of the fact that since f0 primarily affects vowel quality indirectly by affecting the FF-scaling estimate, this effect should approach zero when apparent speaker characteristics are controlled for (since judgments of apparent speaker characteristics are strongly related to FF-scaling). In effect, controlling for apparent speaker characteristics was an attempt to control for listener-internal FF-scaling estimates from trial to trial.

The ASTM should exhibit the same behaviour in that cues that affect vowel quality only by affecting FF-scaling estimates should lose their association to vowel quality when these estimates are controlled for. The ASTM includes an effect for f0 only in triggering changes to FF-scaling estimates, so this effect should essentially disappear when FF-scaling is controlled for. In addition, the experimental design was such that F3+ was strongly tied to FF-scaling and not vowel category (for /æ/ and /ʌ/, the two vowels used in Experiment 1) so that this effect may also be primarily indirect in this case.

Since the ASTM has an FF-scaling estimate associated with each trial, this value could be controlled for directly rather than relying on apparent speaker characteristics. To investigate whether the ASTM also exhibits a decreased sensitivity to indirect cues when the frame of reference is controlled for, the final simulation (from among the ten repetitions) was used. Two sets of correlation coefficients were found. In the first, the marginal correlation between the chosen vowel category (represented by a 1 or 0) and the stimulus properties F1, F3+ and f0 were found. For the second set, the partial correlations between vowel category and stimulus properties were found, after controlling for the FF-scaling estimate associated with the trial. Following Chapter 2, these will be referred to as the no-speaker, and fully-controlled models respectively.

A seen in Table 5.4, the explanatory power of f0 and F3+ is dramatically weakened when FF-scaling estimates are controlled for, while the effect for F1-F2

actually increases. Admittedly, this is not surprising given that f0 affects vowel quality in the PSTM solely by influencing FF-scaling estimates, and the experimental design was such that F3+ was strongly tied to FF-scaling and not vowel category. However, alternative models of vowel perception which include f0 directly in the specification of vowel quality could not generate similar patterns of results, nor could any model that does not have any role for f0. The only sorts of models that can account for the patterns of results generated in this section are those with an indirect role for f0 on perceived vowel quality.

	Observed		
	F1	**F3+**	**f0**
No-speaker Mean	0.824	-0.232	-0.144
Fully-controlled Mean	0.802	-0.215	-0.052
Change in Magnitude	-2.6%	-7.5%	-63.3%

	Simulated		
	F1	**F3+**	**f0**
No-speaker Mean	0.630	-0.192	-0.084
Fully-controlled Mean	0.733	-0.093	-0.000
Change in Magnitude	+8.1%	-48.4%	-99.9%

Table 5.4. *Mean partial correlation coefficients across all 19 participants for the fully-controlled and no-speaker models observed for Experiment 1 (originally presented as Table 2.4) are compared to simulated partial correlation coefficients. The percent change in mean indicates the change in magnitude from the fully-controlled model to the no-speaker model as a function of the magnitude of the no-speaker model.*

In chapter two, all stimulus properties were controlled for in the partial correlation analysis, in addition to apparent speaker characteristics. When conducting this analysis, it was discovered that controlling for both FF-scaling and F1 results in a *positive* partial correlation between f0 and the vowel quality predicted by the model, in contrary to the expected negative relationship. Additional simulations carried out using an unmodified Method 6 of the Sliding

Template Model indicate that this sign flip results from properties inherent to Method 6, and not simply from the modifications made by the ASTM. At the moment, this sign change cannot be explained. The search for a resolution to this issue will be the focus of future research.

5.3.2 Experiment 2

A matrix was created describing the experimental stimuli used in Experiment 2, in the same way it was created for Experiment 1. Three columns contained information specifying the first three FFs, and a fourth column specified the f0 of that vowel. A fifth column contained information regarding changes in participant or changes in block. Each row contained information describing the FFs and f0 for the vowel sound associated with a single experimental trial.

5.3.2.1 Hit Rates

In Experiment 2, listeners were asked to monitor for a single target vowel and to ignore any other vowel they heard. To reflect this, the ASTM considered all ten vowel categories of Edmonton English and not only those explicitly involved in the stimulus design. Hits occurred when listeners correctly indicated having heard the target vowel. Once again, ten simulations of the data were run and the average hit rate was found for each trial across all repetitions for the simulation. This resulted in an average hit rate for each trial. As in Experiment 2, results were organized in terms of voice-pair types based on the acoustic differences between voices in a block. In Figure 5.7, the results of simulations using the ASTM (Simulation A), are compared to those observed in Experiment 2. As seen in this Figure, the ASTM generates a very similar pattern to that of the observed results, including the interaction between FF-scaling differences between voices (referred to as formant-space differences in Experiment 2) and f0 differences on hit-rates.

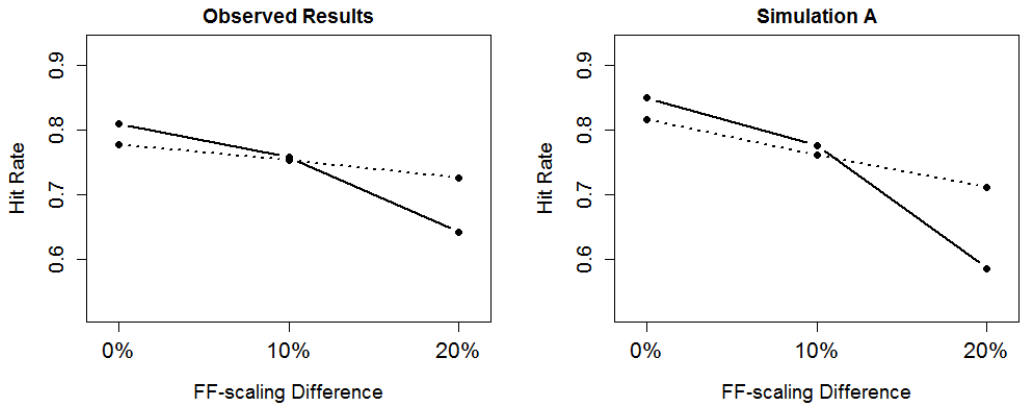

Figure 5.7. *Hit rates are compared for observed results, and those predicted by the full Active Sliding Template Model (Simulation A). The solid line indicates blocks where voices had dissimilar source characteristics, the dotted line indicates blocks where voices had dissimilar source characteristics.*

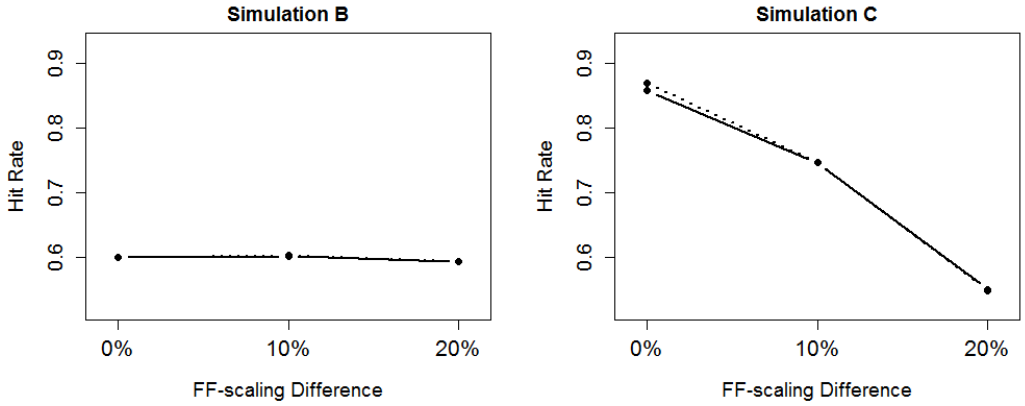

Figure 5.8. *Hit rates are compared for two modified versions of the Active Sliding Template Model. The solid line indicates blocks in which voices had dissimilar source characteristics, the dotted line indicates blocks where voices had dissimilar source characteristics.*

The key to generating this pattern lies in the association between f0 differences and perceived speaker changes. To demonstrate this, two additional simulations were carried out using the same methodology previously outlined. In the second (Simulation B), each new stimulus was assumed to come from a new speaker. This method is essentially Method 6 of the original Sliding Template Model, except for the addition of noise to FF-scaling estimates, and can be

considered a pure-intrinsic model of vowel perception. As seen in the left panel of Figure 5.8, when FF-scaling estimates are not refined, performance is generally low since the classifier does not reduce FF-scaling estimation error by refining estimates using new information in cases where it is appropriate. Furthermore, since only intrinsic information is considered, there is no variation in hit rates based on voice-pair type, since this only affects the kind of extrinsic information present in a block.

In the third simulation (Simulation C), speaker changes were only triggered by changes in block, so that a running FF-scaling was kept regardless of possible changes in speaker. In Chapter 3, normalization methods of this kind were termed passive-extrinsic since extrinsic information was accumulated over a listening situation with little or no active-cognitive control over the organization of this information. Although these models can help reduce estimation error, also included in this simulation, they can inappropriately combine extrinsic information from multiple voices. As seen in the right panel of Figure 5.8, this leads to performance that is negatively related to the formant-space difference (i.e., FF-scaling difference) between voices, with no role for f0 differences between voices.

As outlined in the conclusion of Chapter 3, the pattern presented in the right panel of Figure 5.7 can be thought of as a combination of both panels of Figure 5.8, with Simulation B representing situations where the listener is likely to detect speaker changes (broken line, Figure 5.7), and Simulation C representing situations where the listener is unlikely to detect speaker changes (solid line, Figure 5.7). This pattern arises naturally from a classification system that modifies its behaviour based on the detection of speaker changes by varying from a pure-intrinsic mode to a guided-extrinsic mode of estimating FF-scaling.

5.3.2.2 Reaction Times

In Experiment 2, listeners took longer to respond in blocks made up of vowels from two different voices. Longer reaction times are frequently reported for mixed-speaker listening conditions over single-speaker listening conditions (Summerfield & Haggard, 1973; Mullennix, Pisoni, & Martin, 1989; Magnuson &

Nusbaum, 2007). According to contextual tuning, these increased reaction times are attributable to processes related to normalization which operate unless the reference space has become stable. A stable mapping can be achieved with less effort in single-speaker conditions relative to mixed-speaker conditions, and so reaction times are expected to be shorter overall in single-speaker conditions.

This explanation of events assumes that establishing the reference space is computationally expensive compared to simply classifying a vowel once a mapping has been established. The ASTM shares this characteristic in that the estimation of an optimal FF-scaling for each vowel category involves the most computation in a given trial, while selecting a winning category given an FF-scaling estimate is trivial. Furthermore, as outlined in Section 5.2.2.2, the ASTM monitors for speaker changes, in part, by estimating the best possible match between the vowel sound to be classified and the reference patterns of the current reference space. This 'best match' can then be selected as the winning vowel category in cases where there is no change to the reference space. Since this is carried out by the ASTM for every trial, the instructions involved in the classification of vowel sounds when using an unchanged reference space represent a subset of those involved in the classification of vowel sounds when the reference space is modified.

An additional simulation of the data from Experiment 2 was run, using the same methodology described above. The *microbenchmark* package (Mersmann, 2013) for R was used to determine the amount of processing time devoted to the simulation of each individual trial. This processing time, reported in nanoseconds, will be referred to as CPU time[26]. The average amount of processing time devoted to each trial by the ASTM is compared to observed response times for human listeners in Figure 5.6.

[26] The relationship between processing time and real time cannot be precisely determined using the methods at my disposal. However, the purpose of finding processing times for different trials was only to make gross comparisons of average computational costs in different listening situations.

As seen in Figure 5.9, the ASTM shows similar increases in processing times when voices in a block had different source characteristics. Furthermore, although the pattern is not identical, both observed and CPU response times are positively related to FF-scaling differences between voices. The CPU response time pattern of the ASTM reflects the fact f0 and FF-scaling differences between voices in a block resulted in an increased probability that computationally expensive processes would be involved in a trial by triggering detected speaker changes or decreasing the appropriateness of the extrinsic FF-scaling estimate.

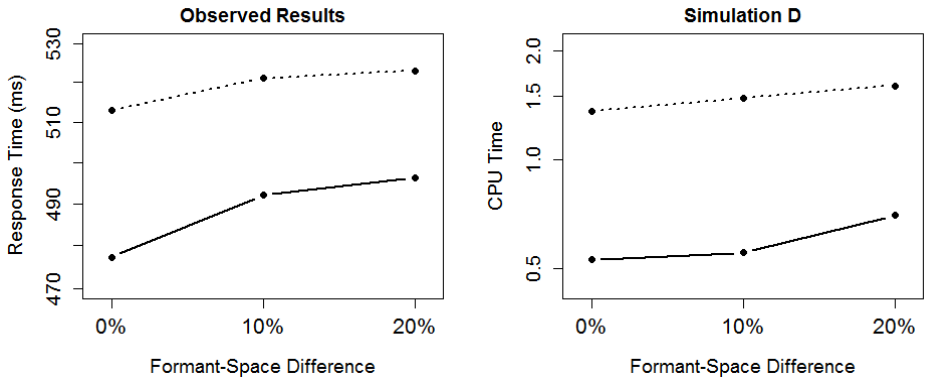

Figure 5.9. *Response times observed for participants in Experiment 2 are compared to CPU times for different voice-pair types. In both cases only times for hit rates are reported. CPU times are in nanoseconds estimated with reference to the CPU clock. Solid lines indicate blocks where voiced had the same source characteristics while the broken line indicates blocks where these differed.*

This pattern of results may be contrasted with the expected CPU time patterns for the pure-intrinsic and passive-extrinsic implementations of the ASTM. In either case, there should be no variation in CPU response times based on voice-pair type, since the same processes are carried out for each trial, regardless of method of presentation. Consequently, these methods would generate patterns like those seen for Simulation B in Figure 5.8. The reaction time patterns shown in Figure 5.9 highlight the fact that, just as with human listeners, normalization processes in the ASTM are not deterministically tied to the presence of multiple voices, but to whether the classifier *acts* as if the listening

situation contains multiple voices. Otherwise, we should expect stable reaction time averages for all voice-pair types other than the single voice (i.e., same source and FF-scaling) condition.

5.4 Conclusion

In Section 5.1, a summary of the experiments contained in this thesis was presented. This interpretation relied on a general frame of reference theory of vowel perception, where the process of normalization is guided by cognitively-active mechanisms, and organized around the detection of speaker changes. In section 5.2, an explicit model of vowel perception that takes into account the insights arising from these experimental results was outlined, and this model was used to simulate the results of Experiments 1 and 2. The results of these simulations indicate that a model of this kind is able to generate the same kinds of patterns of results observed for human listeners, while alternative views of vowel perception without a cognitively-active element are not able to do so. This alignment between theoretical expectations and observed results is taken as a strong indication of the fact that the cognitively-active speaker normalization process outlined above is generally in line with the process of human vowel normalization.

In short, the Active Sliding Template Model presented here can plausibly account for the fact that a) Listeners respond faster and more accurately when presented with vowels from a single voice, b) Listeners can cope well with arbitrary changes in speaker, though this latter condition requires more resources (resulting in increased response times), and is not quite as accurate as when a stable extrinsic estimate of FF-scaling is available, d) Increased processing times associated with mixed-speaker listening conditions are associated with the detection of speaker changes and are not associated with mixed-speaker listening conditions *per se*, and c) Cues that have a primarily indirect effect on vowel quality lose much of their strength when the frame of reference is controlled for. Consequently, it seems reasonable to hypothesize that vowel normalization by human listeners may have at least a grossly similar structure.

Works Cited

Barreda, S. and T. Nearey. (To Appear). The perception of formant-frequency range is affected by veridical and judged fundamental frequency. Proceedings of the 21st International conference on Acoustics, Montreal, 2013.

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, *71*(5), 1150–1166. doi:10.3758/APP.71.5.1150

Irino, T., & Patterson, R. D. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilised wavelet-Mellin transform. *Speech Commun.*, *36*(3), 181–203. doi:10.1016/S0167-6393(00)00085-6

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *The Journal of the Acoustical Society of America*, *88*(2), 642–654. doi:10.1121/1.399767

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. doi:10.1037/0096-1523.33.2.391

Mersmann, O. (2013). microbenchmark: Sub microsecond accurate timing functions. R package version 1.3-0.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. doi:10.1121/1.397688

Nearey, T. M. (1978). Phonetic Feature Systems for Vowels. PhD thesis, Indiana University Linguistics Club.

Nearey, Terrance M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. doi:10.1121/1.397861

Nearey, T. M. and Assmann, P. F. (2007). Probabilistic "sliding template" models for indirect vowel normalization. In Maria-Josep Solé, Patrice Beddor, and

Manjari Ohala (eds.) Experimental Approaches to Phonology. Oxford: Oxford University Press. 246-69.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, speech production, and linguistic structure* (pp. 113–134). Tokyo: OHM.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America*, *117*(1), 305. doi:10.1121/1.1828637

Summerfield, Q., & Haggard, M. P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report of speech research in progress, 2*, 12–23.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100. doi:10.1121/1.393381

Thomson, R. I. (2007). *Modeling L1/L2 interactions in the perception and production of English vowels by Mandarin L1 speakers: A training study*. PhD dissertation, University of Alberta.

Weenink, D. J. M. (2006). *Speaker-adaptive vowel identification*. Doctoral Dissertation, University of Amsterdam. Retrieved from http://dare.uva.nl/document/37721

Appendix 1

Appendix to Experiment 1

The negative partial correlation observed (Section 2.3.1, Table 2.3) between Maleness and Speaker Size judgments is at first glance rather puzzling. However, on further investigation it is clear that there are reasonable explanations for this, which do not affect the interpretation of the other relationships found.

One possible explanation relates to how the Speaker Size ratings were used by listeners. There are two ways that immediately spring to mind: First, absolutely across genders; and second, relatively within genders. In the absolute usage, listeners may have used a single scale, roughly proportional to overall speaker body length (or body mass or volume). In this case, the negative correlation between gender and size judgments would be difficult to explain without bringing further evidence to bear. But in the relative, within-gender usage, a negative partial correlation might readily result. For example, suppose a listener decides a stimulus was an /æ/ that sounded like it was spoken by an individual who was about 165 cm in height, but whose gender was not immediately obvious. It the listener decided ultimately it was a male, they might choose a relatively small size rating because 165 cm is fairly short for a male. However if the listener decided it was a female, they might choose a relatively large size rating, because 165 cm is moderately tall for a female. Suppose on a second replication, the listener made the same assessment of the stimuli, but decided the opposite gender. Cases such as this would contribute to a negative correlation between Maleness and Speaker Size judgments after controlling for all the stimulus factors and vowel judgment.

Another possible explanation involves consideration of the synthetic stimuli in relation to the distribution of acoustic properties measured from natural speech within and across genders. We focus here on f0, which appears to be the

strongest determinant of perceived Speaker Size and Maleness (see Section 2.3.1). The distribution of Speaker-Size responses with respect to the f0 levels used in this experiment will be discussed in reference to data collected by Hillenbrand et al. (1995; vowel data available from http://homepages.wmich.edu/~hillenbr/). This data set consisted of vowels produced by 50 adult males and females, 29 male children and 21 female children (all children were between 10-12 years old). Figure A1.1 presents the distribution of f0s in this data divided by speaker type, while Table A1.1 presents the percentage of tokens from each distribution that exceed the f0 levels used for stimuli in this experiment.

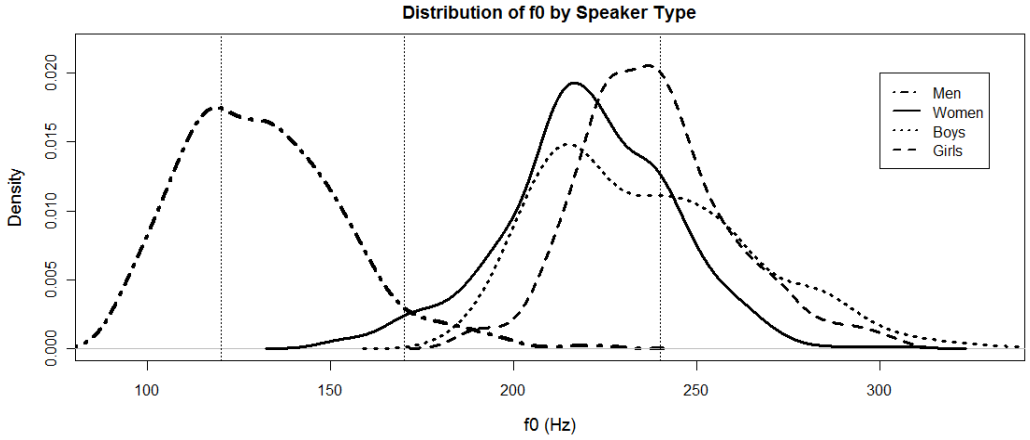

Figure A1.1. *Kernel density plots for the f0 measurements in the data of Hillenbrand et al. 1995. The vertical lines represent the three f0 levels used in the current experiment.*

	Male Adult	**Female Adult**	**Male Child**	**Female Child**
High f0 (240 Hz)	0%	18.6%	40.1%	40.4%
Mid f0 (170 Hz)	5.4%	97.4%	100%	100%
Low f0 (120 Hz)	64.3%	100%	100%	100%

Table A1.1. *Percentage of individual vowels (within each speaker group) from the individual data of Hillenbrand et al. (1995) that have f0 values exceeding the frequencies used in the current experiment.*

Although no adult males in the Hillenbrand data have an f0 as high as 240 Hz, 40.1% of male children's vowels are at least this high. This means that throughout the course of their lives, male speakers have f0s that change from values near those of the high f0 condition to values near those of the low f0 condition. Presumably, at some point during this change they may also have speaking f0s near the mid f0 condition (since this lies between the low and high f0 levels). This naturally leads to a condition in which the f0 levels can be judged as appropriate for a wide range of male speakers, from large to small.

On the other hand, the high f0 level used is close to the average adult female speaking f0 in the Hillenbrand data. As a result, a female speaker with an f0 of 240 Hz may be interpreted as being near normal adult size. The speaking f0 of a typical female speaker does not drop as far as the mid f0 level and would certainly not reach the lowest f0 level. Given that lower f0s are typically associated with larger speakers, vowels with low and mid f0 levels that were interpreted as coming from a female speaker may have led to the impression that the speaker was much larger than the average adult female. The net result of this is that, for any given f0 level, a perceived male speaker will be judged to be smaller than a perceived female speaker (relative to average for that gender).

These facts are reflected in the distribution of Speaker Size responses when grouped by f0 level and gender response as is shown in Figure A1.2. A low f0 level led to the perception of a slightly above average (over all responses) male. Increases in f0 levels lead to movement of the mass of the distribution towards the lower end of the scale, so that Speaker Size responses shifted from slightly over the middle to the bottom of the scale. However, when listeners reported hearing a female speaker, the shift in size responses was much more limited. In the rare cases where listeners heard a female speaker with a low f0, the speaker was reported as very large, usually near the very top of the size scale. As f0 levels increase, the size responses for perceived female speaker also move down the scale, but they settle somewhere around the middle rather than towards the lowest extreme.

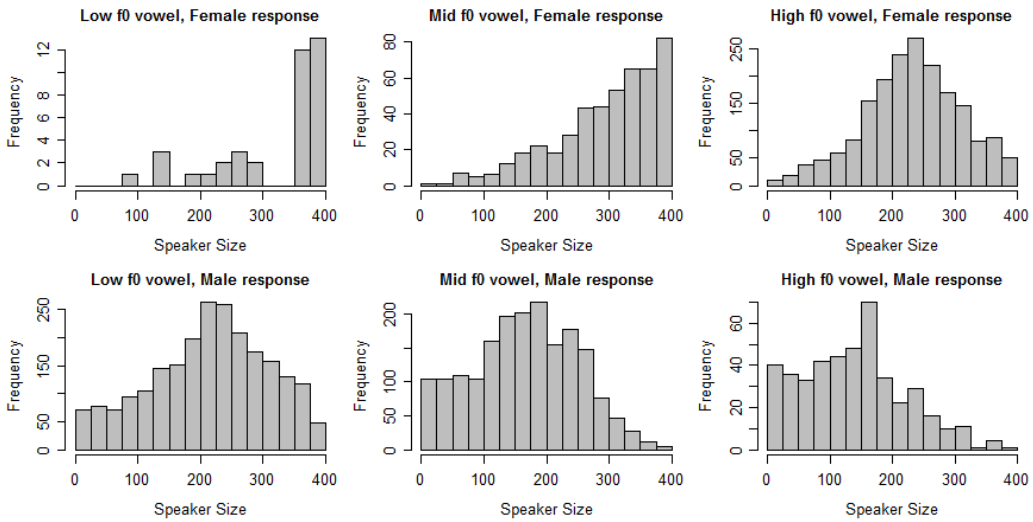

Figure A1.2. *Distribution of speaker size responses grouped by the vowel's gender response and that vowel's f0. Note that each panel has a different y-axis range.*

The relationship exhibited in these graphs is consistent with a negative partial correlation between Maleness and Speaker Size. The within-participant partial correlation was calculated between Speaker Size and Maleness after controlling for f0 only. The average partial correlation was -0.437 ($t = -9.78$, $df = 18$, $p < 0.00001$), which is very similar to the -0.475 value reported for the partial correlation between Speaker Size and Maleness controlling for all other factors (reported in Section 2.3.1). This indicates that the association between perceived Maleness and a (relatively) smaller perceived speaker remains after controlling for the rest of the variables considered in our analysis (F1, F3+, vowel openness).

Works Cited

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97: 3099-3111.

Appendix 2

Appendix to Experiment 2

The source differences between voices in a block were intended to result in the detection of speaker changes. To confirm this, the final 14 participants performed an additional task at the end of each block. Although these participants were not randomly interspersed among all participants, they still represent a random sample of participants in that they were not selected because of any particular quality they possessed. It is important to note that this additional task was not meant to establish a firm connection between acoustic differences between voices in a block and the detection of speaker changes, but only to confirm that, within the context of this experiment, source heterogeneity would strongly signal a likely speaker change. Participants were instructed that at the end of each block they would have to answer two additional questions:

1. How many voices did you hear in the block?
2. How confident are you in that assessment?

At the end of each block, participants were asked to select from two options to answer question one: “one voice” or “more than one voice”. After they answered this question, they were asked to select from the following options to answer question two: “confident” or “unsure”. These options were presented in successive screens so that answering the first question brought up the second question. After answering the second question, participants had a self-timed pause after which they continued on to the next block. Answers to these two questions were analyzed separately as described below.

Since the participants who performed this additional task had their attention explicitly drawn to the number of voices in a block, their performance

may have varied in some way from that of the 51 participants who did not perform the secondary task. To test for this, participants were divided according to whether or not they performed the secondary task, and their hit rates, false alarm rates, and reaction times were sorted according to voice-pair type (as for the analyses presented in Section 3.3.1). A series of independent-sample t-tests was then carried out on hit rates, false alarm rate and reaction times for each voice-pair type, where performance of the secondary task served as the grouping factor. The results of the 18 individual t-tests revealed no significant differences between any of the measurements for any of the voice-pair types, even at an uncorrected p-value of 0.05. The lack of a difference in performance between the two groups may be a result of the fact that, although this secondary task drew explicit attention to the number of voices in a block, it was stated clearly in the instructions given to all participants before commencing the experiment that each block could potentially contain more than one voice, and that this would change from block to block in an unpredictable manner.

A2.1 Number of voices per block

	More than 1 Voice in Block		
	Formant-space Difference		
Voice Source	0%	10%	20%
Homogeneous	4.8 (2.1)	9.8 (4.9)	35.7 (9)
Heterogeneous	97.6 (1.6)	99.1 (0.9)	92.9 (3.1)

Table A2.1. *Percent of rounds in which listeners reported hearing more than one voice in a block, presented by voice-pair type. Numbers in parentheses are the standard errors of each mean.*

The results for this question are presented in Table A2.1. A two-way, repeated-measures analysis of variance was carried out on the rate at which speakers thought a block contained more than one voice. Because of the extreme values for some conditions, an arcsine transform was carried out on the dependent

variable. There were two within-subject factors: the formant-space difference between the two voices (0%, 10%, 20%), and voice source homogeneity. A significant main effect was found for both formant-space difference [$F(2,26) = 6.51$, $p = 0.0051$] and voice source homogeneity [$F(1,13) = 258.53$, $p < 0.0001$], as well as a significant interaction between the two [$F(2,26) = 9.86$, $p = 0.0006$]. When voice sources were heterogeneous, listeners indicated hearing more than one voice in a block in 96.5% of cases, and there is no significant effect for formant-space difference [$F(2,26) = 1.58$, $p = 0.2246$]. When voice sources were homogenous, listeners reported hearing more than one voice in 16.8% of cases and the effect of formant-space difference is significant [$F(2,26) = 11.54$, $p = 0.0040$].

B. Confidence in Number of Voices per Block

	Unsure of Number of Voices in Block		
	Formant-space Difference		
Voice Source	0%	10%	20%
Homogeneous	14.3 (4.4)	25.9 (5)	39.3(5.7)
Heterogeneous	4.8 (2)	4.5 (2.5)	3.6 (2.4)

Table A2.2. *Percent of rounds in which listeners reported being unsure of the number of voices in a block, presented by voice-pair type. Numbers in parentheses are the standard errors of each mean.*

A similar analysis of variance was applied to the rate at which listeners were sure of the number of voices in a block, revealing the same pattern of results, presented in Table A2.2. A significant main effect was found for both formant-space difference [$F(2,26) = 3.72$ $p = 0.038$] and voice source homogeneity [$F(1,13) = 35.78$ $p < 0.0001$], as well as a significant interaction between the two [$F(2,26) = 5.17$ $p = 0.0129$]. When voice sources were heterogeneous, listeners indicated being unsure of the number of voices in the block in only 4.3% of cases and there is no significant effect for formant-space difference [$F(2,26) = 1.58$ $p = 0.2246$].

When voice sources were homogenous, listeners indicated being unsure of the number of voices in the block in 26.5% of cases and the effect of formant-space difference is significant [$F(2,26) = 11.54$ $p = 0.0040$].

C. Summary of Results

When voices in a block had heterogeneous source characteristics, listeners were very likely to hear multiple voices and were confident of this assessment, regardless of the difference in the formant-spaces of the voices. When voices in a block had homogenous source characteristics, listeners were most likely to think that there is a single voice in the block. Even in cases where the formant-spaces of voices differed by 20%, listeners only reported hearing more than one voice in 35.7% of cases. Voice source homogeneity also led to uncertainty regarding the number of voices in the block, and this uncertainty was increased by formant-space differences between voices. Finally, in cases where voices shared source and formant-space characteristics (effectively a single-voice condition), listeners reported being unsure of the number of voices in the block in 14.3% of cases, indicating that the experimental design may have led to a hyper-awareness of speaker-changes.

Appendix 3

Appendix to Experiment 3

It has been suggested that non-uniformities in the vocal tracts of speakers of different sizes might result in the non-uniform scaling of speech sounds between adult males and other speakers (Fant, 1975). Fant suggested that such non-uniformities were due to the relatively longer pharynx-to-mouth ratios of adult males. However, no clear demonstration either of the statistical reliability of systematic non-uniformities nor of the perceptual relevance of any such non-uniformities to listeners' identification performance exist in the literature.

Turner et al. (2009) review difficulties with this hypothesis. In particular, they present a re-examination of the physiological data reported by Fitch and Giedd (1999) and find that although the oral-pharyngeal cavity ratios vary continuously in relation to speaker size, and not simply on the basis of speaker gender, there is no evidence that these differences manifest themselves as differences in produced formant patterns. They conclude that "the anatomical distinction between the oral and pharyngeal divisions of the vocal tract is immaterial to the acoustic result of speech production. For a given vowel, the tongue constriction is simply positioned where it produces the appropriate ratio of front-cavity length to back-cavity length, independent of the location of the oral-pharyngeal junction" (p. 2379). They also state that "speakers adjust the shape of the vocal tract as they grow to maintain a specific pattern of formant frequencies for individual vowels" (p. 2374). Basically, despite differences in anatomy from person to person, speakers strive to produce vowels which differ by a single parameter (i.e., FF-scaling) from the same vowel when produced by other speakers of their language, even if this entails slight modifications to articulatory gestures as a speaker ages.

We do not intend to suggest that vowels vary within-category, between-speakers, solely on the basis of FF-scaling in a deterministic manner. Rather, our position is that, all other things being equal, vowels from speakers of the same dialect different with varying vocal-tract lengths differ in terms of this parameter plus statistical noise. This noise may result from idiosyncratic differences in articulation or speaker anatomy, or it may be a result of the particular situation in which the speech was produced (e.g. clear versus casual speech). The left panel of Figure A3.1 shows the classic Peterson and Barney (1952) vowel data. A visual inspection of Figure A3.1 clearly shows that the major axes of the ellipses are aligned with the F1 = F2 line in a log-space (henceforth ln F1 = ln F2), also indicated on the figure. Variation along the ln F1 = ln F2 indicates equal logarithmic increases to both F1 and F2, and is consistent with variation according to a single multiplicative parameter.

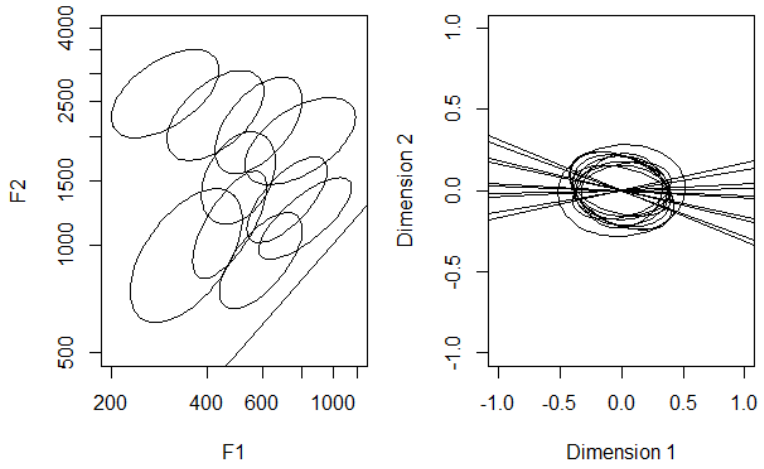

Figure A3.1. *In the left panel, ellipses enclosing two standard deviations of the Peterson and Barney (1952) vowels are presented. The line is a line parallel to ln F1 = ln F2. In the right panel, all formant frequencies have been log-transformed and centered within-category so that vowel-category means are at the origin. All points have been rotated 45 degrees clockwise so that the ln F1 = ln F2 line is now parallel to the x-axis (Dimension 1), while Dimension 2 represents the orthogonal axis. The lines indicate the major axes of the vowel-category ellipses, and they all vary around the ln F1 = ln F2 axis (Dimension 1).*

To investigate the extent of variation along the ln F1 = ln F2 axis, the following analysis was carried out[27]. Formant frequencies were log transformed, and centered according to vowel-category so that all category means were located at the origin. After this, all points were rotated by 45 degrees in a clockwise direction. The result of this is presented in the right panel of Figure A3.1. As a result of these transformations, the x-axis now represents a line parallel to ln F1 = ln F2 and variation along this axis represents variation within vowel-category, between-speakers, that results from uniform logarithmic increases to F1 and F2 (i.e., by a single multiplicative parameter). This analysis revealed that 80.6% of variation between-speaker falls along the ln F1 = ln F2 axis. The same analysis carried out on the vowel data of Hillenbrand et al. (1995) revealed that 79.6% of variation in FFs between speakers falls along the ln F1 = ln F2 axis for that data set. These results are consistent with the hypothesis that variation in FFs within vowel-category, between-speakers, is largely according to a single multiplicative parameter.

[27] This analysis is similar to one presented in Turner et al. (2009). However, that analysis was based on formant wavelengths rather than log-transformed formant-frequencies, which may result in unstable variances. Furthermore, Turner et al. allowed for a specific principal component for each vowel-category ellipse, rather than calculating variation strictly along the axis corresponding to changes in FFs by a single parameter. Allowing for a category-specific slope, and allowing these to vary away from parallelism to the ln F1 = ln F2 line makes that analysis incompatible with a strict uniform scaling hypothesis.

Works Cited

Fant, G. (1975) Non-uniform vowel normalization, STL-QPSR 2-3: 1 – 19.

Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging Journal of the Acoustical Society of America 106: 1511-1522.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America 24: 175-184.

Turner, R. E., Al-Hames, M. A., Smith, D. R. R., Kawahara, H., Irino, T., and Patterson, R. D. (2006). Vowel normalisation: Time-domain processing of the internal dynamics of speech. in Dynamics of Speech Production and Perception, edited by P. Divenyi. Amsterdam: IOS Press. pp. 153-170.